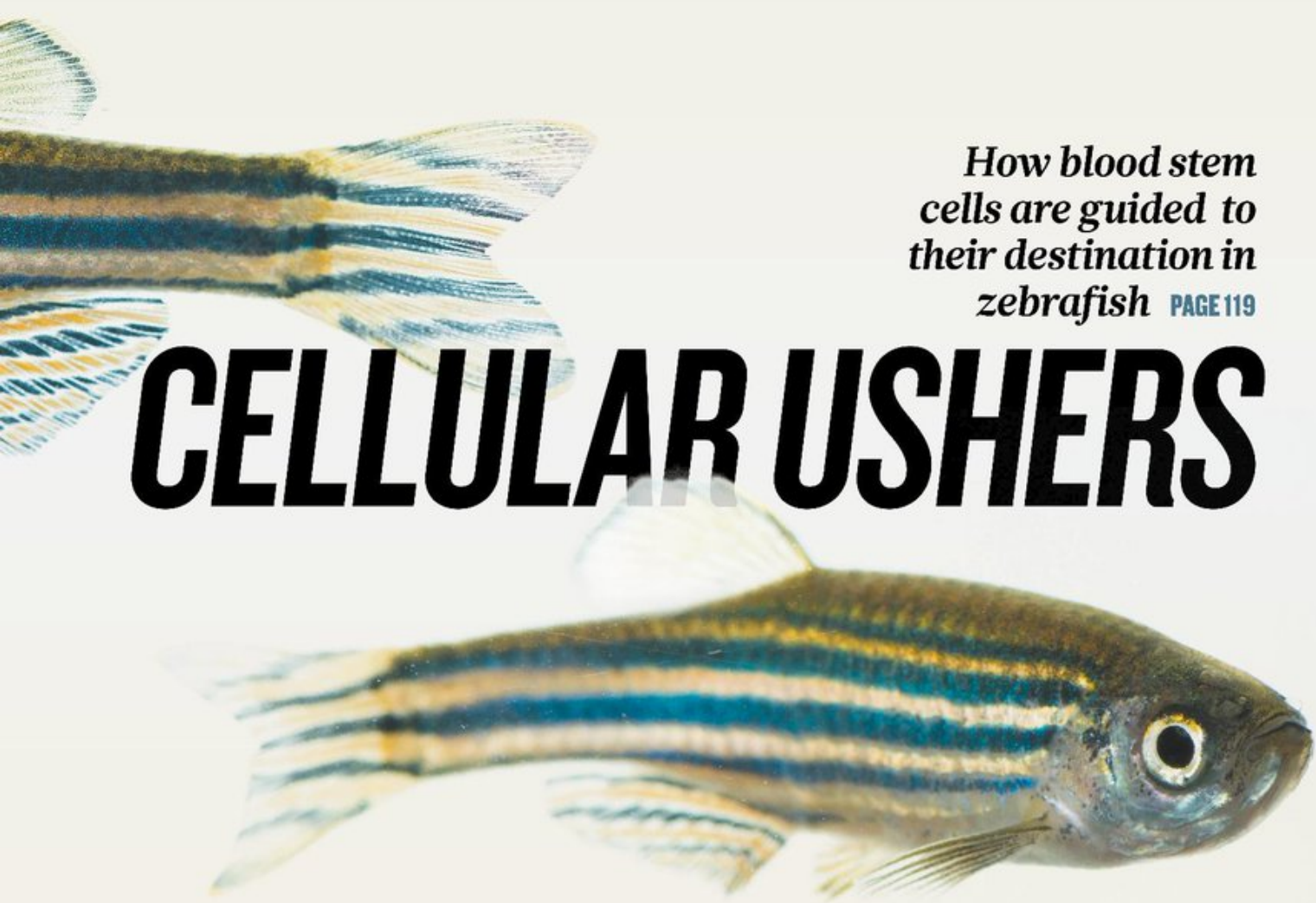


nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



How blood stem
cells are guided to
their destination in
zebrafish **PAGE 119**

CELLULAR USHERS

PLANETARY SCIENCE

WORLD BUILDING

The images reshaping
theories of how planets form

PAGE 20

UN CLIMATE MEETING

TIME TO GET REAL

Zero carbon can't
come soon enough

PAGES 27, 30 & 32

OPTICS

CONTROLLED DEFLECTION

A refractive lens for the
extreme ultraviolet

PAGE 91

NATURE.COM

6 December 2018

Vol. 564, No. 7734

THIS WEEK

EDITORIALS

CLIMATE Annual UN meeting set to agree Paris rules **p.6**

WORLD VIEW Scientists should embrace estimates of certainty **p.7**



FOSSILS Siberian unicorns died out much later than thought **p.9**

How to respond to CRISPR babies

The claims from He Jiankui that he has used gene editing to produce twin girls demand action. A new registry of research is a good start.

People like to say that science is self-correcting. Events in China last week pose a serious challenge to that reassuring platitude. How do researchers respond to the failure of medical ethics, collective responsibility and professional standards that saw an immature experimental technique used to help produce human babies?

It has not yet been independently confirmed that the Chinese genome-editing researcher He Jiankui altered the DNA of embryos using a gene-editing technique and then implanted them in a woman, as he claims. Such a step would be significant and controversial because it would make a permanent change to the germ line that could be passed on to future generations. (This distinguishes germline editing from the use of gene-editing tools as therapies that correct genetic alterations in somatic cells in blood and other tissues.)

Verification of He's claims could be difficult, given that privacy concerns rightly protect the identity of the parents and their one-month-old twin girls. But many scientists in the field agree on two things: the relative simplicity and widespread availability of the gene-editing tool CRISPR-Cas9 mean that what He claims to have done is eminently possible; and, whether or not he is the first person to have genetically edited a baby, he will not be the last.

So, although testing the accuracy of his claim is a priority, so too is ensuring that any future efforts to genetically edit the germ line of human babies proceed in a much more regulated and responsible way. The scientific community still has the opportunity to take the lead on this — public and political reaction to last week's news has been calmer than many might have expected — and it should do so urgently.

Some argue that the circumstances in which germline gene editing would be beneficial, such as to reverse disease-causing mutations that could not be addressed in any other way, are likely to be extremely rare. Nevertheless, given that research and medicine move fast, a clear regulatory system needs to be devised and put in place in case a credible proposal arises. Such a regulatory system should draw on those that already exist to guide the use of gene-editing tools for research into human development, and more broadly govern medical testing of innovative therapies. But it should not start with the assumption that future germline editing is a foregone conclusion — that is a question for society, not scientists, and one that demands the input of different stakeholders from across the world. Researchers and physicians must ask permission rather than beg for forgiveness.

A solid regulatory system set up by the research community can then be the basis for laws and regulations that individual nations might decide to introduce. Debate was key to framing the law that regulates a mitochondrial-replacement therapy in the United Kingdom, a procedure that also affects unborn babies and means they carry DNA from three people. (Laws are not always the best way to govern emerging medical procedures, but they do offer the deterrence of effective punishment for those who don't follow the rules, unlike self-regulation or guidelines.)

So, how can the gene-editing community set up a better system? A

starting point would be a global registry (or national registries) set up by funders or governments to record preclinical research that involves gene editing in human embryos. This would require the objectives, steps and limitations of projects to be spelled out from an early stage. The records should also detail the steps taken for ethical approval and oversight of the research. The 2016 guidelines from the International

"The scientific community still has the opportunity to take the lead."

Society for Stem Cell Research are a good model to follow for regulation of research that involves human embryos and gametes, including research into germline gene editing.

Such registries could also provide a mechanism to flag research projects that do not meet high ethical and technical standards, and a route to apply pressure on individuals and their institutions to improve. And they could provide a framework, if the time comes, to define a path to the clinic. They would help to explain the risks and potential benefits to people — such as prospective parents — so they can make more informed choices.

He's claims to have communicated his intentions and actions to the scientific community do not stand up to serious scrutiny. The community — from individual researchers to institutions — can and must do more to encourage more meaningful, transparent engagement and discussion on specific projects. In return, scientists who are trusted to carry out research have the responsibility to welcome and embrace scrutiny. ■

A lonesome life

Genome of legendary Galapagos giant tortoise shares some secrets of longevity.

Lonesome George, the last member of *Chelonoidis abingdonii*, a species of giant tortoise endemic to the tiny island of Pinta in the Galapagos Islands, did not die in vain. Researchers this week present his genome in the journal *Nature Ecology and Evolution* (V. Quesada *et al.* *Nature Ecol. Evol.* <https://doi.org/10.1038/s41559-018-0733-x>; 2018), along with the genome of George's distant but still-extant cousin, the Aldabra giant tortoise *Aldabrachelys gigantea*. Comparison of these genomes with those of a diverse range of species unlocks a treasure trove of secrets about how giant tortoises get to be so large, long-lived (typically up to a century) and resistant to infections and cancer.

Once upon a time, islands from Malta to Mauritius could boast their own species of giant tortoise. But nowhere is more synonymous with giant tortoises than are the Galapagos Islands — literally so, because the archipelago gets its name from *galápagos*, a Spanish word for turtle.

Marooned in isolated spots and free from predators, Galapagos tortoises became larger than their mainland ancestors, and, having rather relaxed metabolisms, they are able to survive on the meagre rations available on islands. Slow metabolism and large size tends to correlate with long life and infrequent reproduction. It's no surprise, therefore, that the arrival of humans marked out giant tortoises as ripe for extinction. These large creatures moved too slowly to escape slaughter, and bred too infrequently to compensate for the loss. Even when they did manage to breed, their eggs and young were easy prey for other introduced species such as rats, the eradication of which is seen as key to the recovery of giant tortoise populations (see W. T. Aguilera *et al.* *Nature* **517**, 271; 2015).

Humanity, however, wasn't solely to blame. Comparison of the genome of Lonesome George — who died in 2012 — with that of other tortoises shows that the effective population size of his species had been in slow decline for at least one million years. This is only to be expected for a species of large, slowly reproducing animal confined to a small island, where the choice of mate is limited. The Aldabra giant tortoise experienced more ups and downs; but for isolated island species, downs can all too often prove catastrophic.

Animals that live for a long time take pains to avoid early death, and giant tortoises are among the longest-lived of all land animals. Although the genetics of longevity has been explored in long-lived mammals, extending it to tortoises should illuminate more-general hallmarks of the genetic basis of longevity.

Genes under positive selection in giant tortoises include those whose expression has also been connected with a ripe old age in humans. A detailed study of 891 genes involved in the function of the immune system revealed duplications in tortoise genes not seen in humans, and there are more tumour-suppressor genes in giant tortoises than in vertebrates in general. Duplications of at least one proto-oncogene involved in mitochondrial health might relate to

an improved response to oxidative stress, known to be an important factor in ageing.

Some details of the giant-tortoise genomes could shed light on aspects of the peculiar evolution and development of tortoises, such as their shell. One should therefore be cautious in applying the lessons of tortoise longevity directly to humans.

The longevity of a species is more than a matter of a list of genes — it's connected with all aspects of the species' life history. Although the naked mole-rat (*Heterocephalus glaber*) can live for 30 years, this marks it out as peculiarly long-lived only for rodents, whose lives are generally fast, frenetic and short. It's no great shakes compared with a tortoise, a human or indeed a bowhead whale, whose two-century lifespan makes it the longest lived of all mammals — and which doubtless has many other whale-specific peculiarities. Faced with the specific fate of one's species, life remains very much what you make it. ■

“One should be cautious in applying the lessons of tortoise longevity directly to humans.”

Climate rules

Global leaders have gathered to decide on emissions guidelines — but time is running out.

Delegates to the United Nations climate talks arrived in the old Polish coal-mining town of Katowice at the weekend to learn that the annual meeting faces an uncertain future: incoming Brazilian president Jair Bolsonaro has withdrawn his country's offer to hold the event next year. This unwelcome posturing, from a leader who seems likely to oversee renewed deforestation in the Amazon, shows that global warming is far from the top of the political agenda in some countries. But it also acts as a reminder that political cooperation remains the only effective defence we have against the worst effects of climate change — which would mean a more hostile world for us all.

The annual caravan of government representatives, campaigners and negotiators has rolled into Poland for the 24th Conference of the Parties to the United Nations Framework Convention on Climate Change (COP24) with a clear goal. Delegates from more than 190 countries hope to finalize the rules for how the 2015 Paris climate agreement, which aims to limit global warming to no more than 2°C above pre-industrial levels, will be put into practice. Negotiating an acceptable plan for curbing emissions and funding climate action will be a tough task. But given the enormity of the environmental and social challenges ahead, there is a need for more than written rules and good intentions.

The Paris agreement is a hybrid of self-imposed national commitments and binding 'top-down' elements, including mandatory emissions reporting and a regular global stock-take of collective progress. Transparent rules and criteria for cooperation among nations, including systems that link countries' individual actions through international carbon markets, are essential for the success of an agreement otherwise plagued by the voluntary nature of national climate targets.

Despite decades of international climate diplomacy, global greenhouse-gas emissions continue to rise. The concentration of carbon dioxide in the atmosphere is now at a level that Earth hasn't experienced for several million years. Since 1900, global temperatures have already increased by 1°C — with inescapable consequences. Raging forest fires

last month in drought-stricken California are a clear warning sign of what a warmer future might hold in store (see Comment, page 27).

A special report released in October by the Intergovernmental Panel on Climate Change found that time is running out to limit global warming to 1.5°C. Realistically, that horse has already bolted. To keep warming to 2°C — which would still all but guarantee severe environmental effects — global emissions would need to shrink by at least one-quarter by 2030, and drop to almost zero by 2050. But according to a report released last week by the UN Environment Programme, there is a huge gap between nations' self-imposed targets and the amount of action that is needed to stabilize the climate.

In particular, the world's largest greenhouse-gas emitters, including China, the United States and the European Union, must significantly step up their own efforts to tackle climate change. But will they? US President Donald Trump has already said that the United States will pull out of the Paris agreement, claiming it is bad for the economy. But a report issued by 13 federal agencies in November found that the US economy could shrink by as much as 10% by 2100 if little is done to reduce global warming, and several US states and cities have unveiled their own ambitious emissions-reduction pledges.

Whether China will be able and willing to decarbonize its fossil-fuel-based economy in due time is uncertain, despite encouraging signals from the leadership. China's emissions reporting and verification practices are notoriously non-transparent. The Paris rulebook aims to bolster these mechanisms, and China must show its support for this.

The EU seems best placed to take climate policies to a higher level (of ambition, at least). Ahead of the Katowice conference, the European Commission released a set of scenarios for how the bloc can achieve zero net emissions by 2050 — although member states must still agree on the preferred scenario. Poland and other EU countries that rely heavily on coal might oppose more ambitious targets. In Germany, too, the timing and cost of the planned phase-out of coal-powered plants are causing heated debate. But the EU's initiative is a strong signal that the push for clean energy must involve all sectors of the economy, including industry, transport, building and agriculture.

Katowice, a European coal capital, is an apt place to meditate on the future of fossil fuels. Behind the razzmatazz of these climate-policy talks are simple facts: the world's policymakers must introduce more and stronger measures to boost investment in clean energy and end the use of dirty fuels. Delay is fundamentally contrary to reason. ■



How sure are you of your result? Put a number on it

Any scientist publishing a claim should quantify their confidence in it with a probability, argues Steven N. Goodman.

Picture our bafflement if weather forecasts said, “There is a non-statistically significant chance of rain tomorrow,” rather than, “There is a 60% chance of rain.” Or the confusion among Florida residents if, instead of being told, “Tallahassee has an 85% chance of a direct hit by Hurricane Michael,” they heard: “Tallahassee will be hit, $P=0.03$.”

When the stakes are high, we need accurate and understandable risk estimates to make informed decisions. We demand that weather services clearly convey the chance of rain or hurricane, because lives and livelihoods are at stake. That’s why forecasts distil immensely complex models into one number.

Scientists should do the same.

Let’s require that any researcher making a claim in a study accompany it with their estimate of the chance that the claim is true — I call this a confidence index. As well as, “This drug is associated with elevated risk of a heart attack, relative risk (RR) = 2.4, $P=0.03$,” investigators might add: “There is an 80% chance that this drug raises the risk, and a 60% chance that the risk is at least doubled.”

Analyses using Bayesian statistical methods, which generate the probability of a hypothesis being true, go part way down this path. For example, a 2017 study calculated that induced hypothermia has a 76% chance of benefiting newborn babies who have brain damage from oxygen deficiency, even though the test and control groups were not statistically different — I calculated the P value for this study to be 0.6 (A. R. Laptook *et al.* *J. Am. Med. Assoc.* **318**, 1550–1560; 2017).

A confidence index would formally incorporate the impact of previous evidence (as some Bayesian analyses do) and investigators’ judgement about the plausibility of a claim’s explanation. Importantly, a confidence index should capture the limitations of the study that are currently addressed only qualitatively. It would apply whether or not researchers also calculate confidence intervals, a separate metric.

Many scientists assume that the P value is a confidence index; a widespread, mistaken belief is that a P value under 5% implies a 95% or greater probability of the effect. But the P value does not measure the probability that the null hypothesis is true. This stubborn misconception so distressed the American Statistical Association that in 2016 it issued a rare public statement to dispel it, and to discourage the use of ‘bright line’ P -value thresholds (usually 0.05) to justify claims (see go.nature.com/2p9hcxn). More reliance on confidence intervals has been proposed as a remedy, but these are also often used in ‘bright-line’ fashion and share many of the limitations of P values. Most crucially, a confidence interval from a reliable study can be identical to one from a study in which we have zero confidence.

Claims are often communicated so obliquely that it is hard to know what to make of them. If it is statistically significant, the existence of a relationship is asserted as if it is definitively true, as in “Fibre intake reduces cancer risk by 18% (confidence interval 2% to 34%), $P=0.02$ ”. If the result is not statistically significant, an array of statements is possible, from “there is no difference ...” to “there is a trend ...” and various other creative circumlocutions.

This fuzziness makes clear communication difficult between scientists, and all but impossible with others, from journalists to doctors, policymakers and the public.

Some people will say that they cannot translate all the nuances of research into one number, but the practice has ample precedent. The Intergovernmental Panel on Climate Change puts confidence levels on its statements. Crowdsourcing techniques such as prediction markets, in which people place bets on outcomes, have been used to estimate with decent accuracy the chance that scientific studies will be replicated. Some people have proposed that scientists bet their own money on their claims, in part to discourage over- or under-confidence and other cognitive pitfalls.

A confidence index could help in other ways. Some researchers manipulate analyses or selectively report outcomes to achieve statistical significance (a process called P hacking) because publication, recognition and funding are most likely to flow from statistically significant studies. This year, a survey of 390 biostatisticians found that at least 20% had been asked by a collaborator to manipulate their data or analysis to exaggerate their results’ importance (M. Q. Wang *et al.* *Ann.*

Intern. Med. **169**, 554–558; 2018). With a confidence index, because there is no ‘bright line’ to aim for, the incentive to hack it might be replaced by an incentive to get it right.

Of course, the reasoning and method used to calculate a confidence index should be reported. The foundations for such methods already exist — in Bayesian statistics, sensitivity analyses and more — although they need further development. And then there is scientific judgement: the same judgement behind the words currently used. But numbers always speak louder than words, and it is beyond time to convert those words into numbers with clearer meaning.

Although simple on paper, requiring a confidence index would entail a profound overhaul of scientific and statistical practice. But the crisis of reproducibility and credibility in research demands no less. Crises should not be wasted; if there was ever a time for transformation, it is now. ■

Steven N. Goodman is professor of medicine and of epidemiology at Stanford University in California.
e-mail: steve.goodman@stanford.edu

FUZZINESS
MAKES CLEAR
COMMUNICATION
DIFFICULT
BETWEEN SCIENTISTS,
AND ALL BUT
IMPOSSIBLE
WITH OTHERS.

SEVEN DAYS

The news in brief

CLIMATE

El Niño forecast

There's a 75–80% chance that a weak El Niño weather pattern will develop between December 2018 and February 2019, according to the World Meteorological Organization (WMO). Researchers don't expect the upcoming El Niño to wreak as much havoc as did the monster 2015–16 event, which flooded parts of South America and sparked worldwide coral bleaching, among other things. But even a weak El Niño could boost global temperatures and lead to increased rainfall in areas such as the southeast coast of South America, the WMO said on 27 November. Sea surface temperatures in the central and eastern tropical Pacific warmed to weak El Niño conditions in October. Researchers predict that the atmosphere will respond to the increased temperatures with changes in wind and cloud patterns in the coming months.

Drastic action

Governments of the world need to triple their current efforts to reduce greenhouse-gas emissions to prevent global warming of more than 2 °C above pre-industrial levels by 2030, the United Nations Environment Programme (UNEP) said in its annual 'emissions-gap' report. Released on 27 November — just a week before the latest UN climate summit in Katowice, Poland — the report projects that current national policies would allow global greenhouse-gas emissions to rise by around 10% by 2030, compared with 2017 levels. But emissions would need to decrease by 25% over the same period to maintain a probable chance of limiting warming to 2 °C. Nations would need to reduce emissions by 55% to restrict warming to below 1.5 °C.



JOHN WESSELS/AFP/GETTY

Ebola outbreak hits bleak milestone

The ongoing Ebola outbreak in the Democratic Republic of the Congo (DRC) is now the second largest on record, according to the World Health Organization. "This is a milestone nobody wanted to hit," said a spokesperson for the agency in an e-mail to *Nature*. With 444 total cases, including 260 deaths, as of 2 December, the outbreak is larger than the 9 others recorded in the country since 1976. It is still smaller than

the 2014–16 Ebola crisis in West Africa, which resulted in more than 28,000 cases, including about 11,300 deaths. As the DRC outbreak spreads into cities whose inhabitants are living amid conflict, and with refugees regularly moving around the region and into the neighbouring countries of Uganda, South Sudan and Burundi, officials warn that the situation is unpredictable and that the outbreak will continue well into next year.

FACILITIES

Oil surveys

The US National Marine Fisheries Service will allow five companies searching for offshore oil to use a technique to that can harm animals such as whales and dolphins. Air-gun blasts used in these surveys create powerful sound waves — which can be louder than a Saturn V rocket at launch — that help to map oil and gas deposits below the sea floor. The authorizations, issued on 30 November, limit when and where air-guns can be fired along the US east coast. They also include monitoring and reporting requirements to help mitigate the survey's impacts on marine mammals, which are protected under a 1972 law. The authorizations are the first

step of a two-part process as US President Donald Trump seeks to open coastal waters to oil exploration. The Bureau of Ocean Energy Management must now issue permits before the surveys can begin.

POLICY

Gene-drive treaty

Nations rejected a proposal to temporarily ban the release of organisms carrying gene drives — a genetic-engineering technology designed to spread mutations rapidly through a target population — on 29 November at a meeting of the United Nations Convention on Biological Diversity (CBD) in Sharm El-Sheikh, Egypt. Dozens of scientists opposed the moratorium proposal, although numerous

environmental and activist groups supported it. A gene-drive moratorium was never likely to succeed in the face of opposition from biotechnology-friendly countries, because changes to the CBD require a consensus among the convention's almost 200 parties. Instead, representatives at the two-week-long meeting agreed on changes that were vague enough for both proponents and sceptics of gene-drive technology to claim victory. Signatories to the CBD, which has been ratified by most of the world's countries and influences national laws that affect biodiversity, agreed on the need to assess the risks of gene-drive releases on a case-by-case basis. They also said that local communities and

NASA/GODDARD/UNIV. ARIZONA
Indigenous groups potentially affected by such a release should be consulted.

PEOPLE

French connections

A controversy over the leadership of France's national biomedical-research agency INSERM ended on 26 November with the news that the government had appointed Gilles Bloch, president of the University Paris-Saclay. Bloch succeeds Yves Lévy, who had sought a second four-year term as INSERM chief, but pulled out for "personal reasons". Lévy's leadership, which began in 2014, had become controversial after his wife, Agnès Buzyn, was appointed health minister in 2017 — although procedures were put in place to mitigate a possible conflict of interest between the two. Bloch will take up the post on 2 January.

Science minister

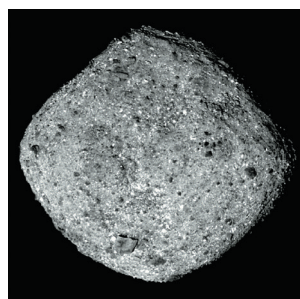
Britain's universities and science minister, Sam Gyimah, resigned on 30 November after Prime Minister Theresa May said that the country would not seek to continue using Galileo, the European Union's satellite-navigation system, after Brexit. Britain's future participation in Galileo has been a sticking point in

Brexit negotiations: the UK government had intended to negotiate rejoining the system, but EU law dictates that a non-member state cannot be involved in developing the secure part of the system, which provides signals for government users, including the military. Gyimah also said that he would vote against the Brexit divorce deal — which defines the terms of the country's exit — when it comes before Parliament on 11 December. The deal, agreed last month between UK and EU officials, has divided politicians and prompted a spate of ministerial resignations. Gyimah became science minister in January. His replacement had not been named as *Nature* went to press.

SPACE

Asteroid arrival

NASA's OSIRIS-REx spacecraft arrived at its target, the 500-metre-wide asteroid Bennu, on 3 December after a nearly 27-month journey. The probe will now loop around Bennu (pictured) to study it until 2020. Then, in July of that year, OSIRIS-REx will lower itself to the asteroid's surface, pick up at least 60 grams of asteroid dirt and fly the sample back to Earth. It is scheduled to return in 2023.



OSIRIS-REx — or Origins, Spectral Interpretation, Resource Identification, and Security–Regolith Explorer — also aims to study the factors that affect the paths of potentially hazardous asteroids. It will be the first US mission to return to Earth after collecting an asteroid sample.

Black-hole bounty

Astronomers have announced the detection of four new gravitational-wave events — ripples in the fabric of space-time created by cataclysmic cosmic events. The signals, detected in 2017, were created by mergers of black holes, and include hints of the largest such merger yet, which produced a black hole more than 80 times as massive as the Sun. The studies were posted on the website of the US-based Laser Interferometer Gravitational-Wave Observatory (LIGO) collaboration. The experiment — which announced its

first historic detection of gravitational waves, from a black-hole merger, in 2016 — has now detected ten such events, as well as one collision of two neutron stars that produced the strongest gravitational-wave signal yet.

PUBLISHING

Open-access push

More than 1,400 researchers have signed an online letter supporting Plan S, an initiative backed by 16 research funders that mandates that, by 2020, papers resulting from their funding should be free to read instantly on publication. The petition, launched on 28 November by Michael Eisen, a geneticist at the University of California, Berkeley, comes as scientists continue to debate the European-led plan. A letter published 3 weeks earlier — which has more than 1,400 signatures — had called Plan S "a serious violation of academic freedom", because it would prevent researchers from publishing where they wanted. But Eisen's petition argues against this, and says that although funder mandates might "superficially limit our publishing options in the short term", they would ultimately lead to a system that maximizes the reach of scholarship and its value to all.

TREND WATCH

On 8 December, Chang'e-4 will set off to become the first spacecraft to touch down safely on the Moon's far side. Half a century after the original space race, there has been a resurgence of interest in Earth's satellite, with some scientists saying we are entering a renaissance of Moon exploration.

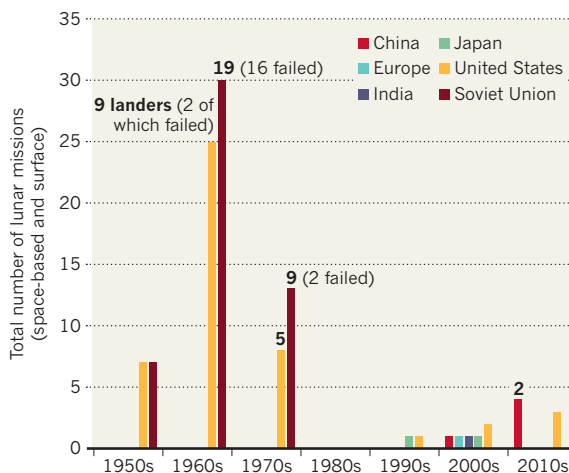
The United States and the Soviet Union conducted a large number of lunar missions — many unsuccessful — in a race to prove dominance in space exploration. These culminated with the Apollo missions that put humans on the Moon in 1969. After the excitement of the space

race, it went quiet. The Soviet Union sent its final mission to the Moon in 1976, and budgetary issues and a lack of political will in the United States led to missions being cancelled.

But the past two decades have seen a gradual return of lunar missions. Some dozen of these orbiters have been launched since 1990, by the United States, but also by new players such as China and Europe. Landers are also back in fashion. In 2013, Chang'e-3 became the first mission to land on the Moon since the 1970s. Chang'e-4 will follow, and landers from Japan and Russia are also planned.

REACHING FOR THE MOON

Humanity's interest in sending spacecraft to the Moon peaked in the 1960s, reached a low in the 1980s and is now bouncing back — with countries such as China joining the space race.

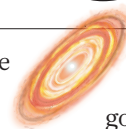


NEWS IN FOCUS

ENVIRONMENT Plan to save Venice from floods threatens lagoon ecosystem **p.16**

IRAN Extent of Tehran's sinking laid bare in satellite images **p.17**

CONFLICT Geologists measure bullet damage to ancient settlements **p.18**



ASTRONOMY The rulebook for world-building just got more complicated **p.20**

MARK SCHIEFELBEIN/AP/SHUTTERSTOCK



The CRISPR-Cas9 tool was used to genetically modify embryos before implanting them into a woman.

GENE-EDITING

CRISPR-baby scientist fails to satisfy his critics

He Jiankui reveals details of how he edited babies' genomes, but many questions remain.

BY DAVID CYRANOSKI

He Jiankui, a Chinese scientist who claims he helped to produce the first people born with edited genomes — twin baby girls — appeared at a gene-editing summit in Hong Kong to explain his experiment. On 28 November, he delivered his talk amid legal threats and mounting questions about the ethics of his work, which until then he had outlined largely in YouTube videos.

Scientists welcomed the fact that he

appeared at all — but his talk left many of them hungry for answers, including whether his claims are accurate.

“There’s no reason not to believe him,” says Robin Lovell-Badge, a developmental biologist at the Francis Crick Institute in London. “I’m just not completely convinced.” An independent body should confirm the test results by thoroughly comparing the parents’ and children’s genes, say Lovell-Badge and others.

Many scientists faulted He for the seemingly cavalier nature in which he embarked on such

a landmark, and potentially risky, project.

“I’m happy he came, but I was really horrified and stunned when he described the process he used,” says Jennifer Doudna, a biochemist at the University of California, Berkeley, and a pioneer of the CRISPR-Cas9 gene-editing technique that He used. “It was so inappropriate on so many levels.”

Alta Charo, a bioethicist at the University of Wisconsin–Madison and a member of the summit’s organizing committee, says: “Having listened to Dr He, I can only conclude ▶

► that this was misguided, premature, unnecessary and largely useless.”

CCR5, the gene that He edited using CRISPR-Cas9, is the door through which many strains of HIV infect immune cells. Many scientists have criticized He's choice to alter this gene, in part because there are other ways to stop people from contracting HIV. Critics also say that other diseases would make more obvious targets for elimination through editing embryonic genomes. Huntington's disease or Tay-Sachs disease are examples of conditions that, in some circumstances, might be averted only through gene editing, said George Daley, dean of Harvard Medical School in Boston, Massachusetts.

HIV-RESISTANT TWIN

He revealed that one of the genetically modified twins will be resistant to HIV, because the gene edits removed both copies of her CCR5 gene. The other twin could still be susceptible to infection, because the gene-editing process inadvertently left one of her copies of CCR5 intact, he said.

He's decision to implant the second embryo drew strong criticism. “Why choose this embryo? It just doesn't make sense scientifically,” said geneticist Jin-Soo Kim of Seoul National University. He Jiankui said he had explained the situation to the parents and they decided they wanted to do it anyway. He also made clear that his aim is to prepare the technique for global use: “For millions of families with inherited disease or infectious disease, if we have this technology, we can help them.”

He initially worked with eight couples in which the men were HIV-positive and the women HIV-negative, but one couple later dropped out of the study. His team first washed the men's sperm to ensure that HIV was not present. The researchers then injected the sperm, and CRISPR-Cas9 enzymes, into unfertilized eggs from the men's partners. This produced 22

embryos, of which 16 seemed to be viable and to have been edited. Two of the four embryos from one couple contained modifications to CCR5, and He says that he implanted these, even though one embryo still had an intact copy of the CCR5 gene, to produce the twins.

It is not clear what has happened to the other embryos. He said that he has now put the experiments on hold, but that he had already implanted a gene-edited embryo into another woman.

Kim says he's 90% sure that He succeeded in editing the twins' genomes as claimed, in part because He used state-of-the-art sequencing methods before and after implantation to show that the embryos contained no unwanted mutations.

But He's talk leaves a host of questions unanswered, including whether the prospective parents were properly informed of the risks; why He selected CCR5 modification when there are other, proven methods for HIV prevention; why he chose to do the experiment with couples in which the men have HIV, given that women with HIV have a higher chance of passing the virus on to their children; and whether the risks of knocking out CCR5 — which could have necessary but still unknown functions — outweighed the benefits.

Nobel-prizewinning biologist David Baltimore, chair of the summit's organizing committee and former president of the California Institute of Technology in Pasadena, called He's experiment “irresponsible”. Baltimore also accepted blame on behalf of the scientific community: “There has been a failure of self-regulation.”

In response to questions about why the community had not been informed of

the experiments before the women were impregnated, He cited presentations he gave last year at meetings at the University of California, Berkeley, and the Cold Spring Harbor Laboratory in New York. But Doudna, who organized the Berkeley meeting, says that He did not present anything showing he was ready to experiment in people.

He also said that he discussed the human experiment with unnamed scientists in the United States. But Matthew Porteus, who researches gene-editing at Stanford University in California, says that's not enough for such an extraordinary experiment. Porteus wants He to post his data to a server such as bioRxiv, so that other scientists can analyse them.

BIG EXPECTATIONS

Pressure was mounting on He ahead of the presentation. On 26 November, the Chinese national health commission requested the Guangdong health commission — which is in the same province as He's university — to investigate. The Chinese Academy of Sciences has issued a statement condemning He's work, and the Genetics Society of China and the Chinese Society for Stem Cell Research jointly issued a statement saying that the experiment “violates internationally accepted ethical principles regulating human experimentation and human rights law”.

The hospital cited in China's clinical-trial registry as having given ethical approval for He's work posted a press release on 27 November saying it did no such thing. The hospital, itself now under investigation by health authorities, questioned the signatures on the approval form and said that its medical-ethics committee had never held a meeting related to He's research. He has not responded to *Nature's* requests for comment on these statements and investigations.

On 28 November, Francis Collins, director of the US National Institutes of Health (NIH), said in a statement that “this work represents a deeply disturbing willingness by Dr He and his team to flout international ethical norms”.

Fears are growing that He's actions could stall the responsible development of germline gene-editing, the modification of genes that are passed on to future generations. At the summit, Daley urged support for such research: “It's possible that the first instance came forward as a misstep, but that should not lead us to stick our heads in sand and not consider a more responsible pathway to clinical translation.”

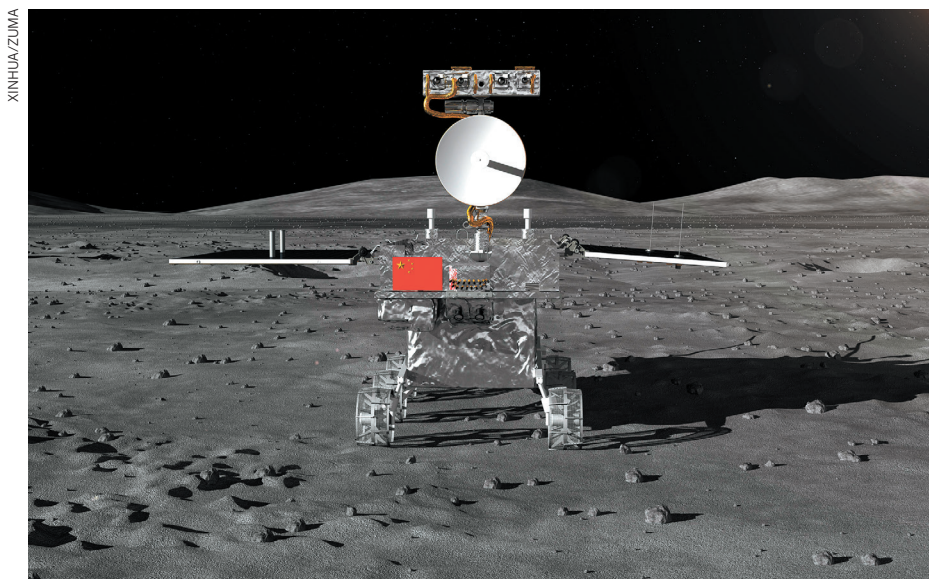
The pressures facing He were clear ahead of his talk, in particular when Lovell-Badge made a plea uncharacteristic of scientific meetings. “He has to be given a chance to explain what he did,” said Lovell-Badge. “We cannot have unruly behaviour.” There was also heightened security, with men in dark suits near the stage, and cameras lining the back of the auditorium. Porteus says that He's appearance was a first step, but that He will have to start answering lingering questions soon. “He's already at risk of becoming a pariah.” ■

“He has to be given a chance to explain what he did. We cannot have unruly behaviour.”



He Jiankui faced tough questions after his talk at the gene-editing summit in Hong Kong.

TPG VIA ZUMA



That's one small step for a rover, one giant leap for lunar science.

MOON EXPLORATION

Journey to the far side of the Moon

China is about to land on the Moon's dark side; experiments will include vegetable growing and radio astronomy.

BY ANDREW SILVER

Early in the New Year, if all goes well, the Chinese spacecraft Chang'e-4 will arrive where no craft has landed safely before: the far side of the Moon. The mission is scheduled to launch from Xichang Satellite Launch Centre in Sichuan province on 8 December.

The craft, comprising a lander and a rover, will then enter the Moon's orbit, before making a controlled landing on the surface. The mission's main job will be to investigate this side of the lunar surface, which is peppered with many small craters. The lander will also conduct the first radio-astronomy experiments from the far side of the Moon — and the first investigations to see whether plants will grow in the low-gravity lunar environment.

"This mission is definitely a significant and important accomplishment in lunar exploration," says Carolyn van der Bogert, a planetary geologist at Westfälische Wilhelms University in Münster, Germany.

The ultimate goal of the mission is to create a Moon base for future human exploration. Chang'e-4 will be the country's second craft to 'soft' land on the lunar surface, following Chang'e-3's touchdown in 2013.

The China National Space Administration

(CNSA) has remained tight-lipped about many of the mission's details, including the landing site. The most likely location is inside a 186-kilometre-wide crater called Von Kármán, says Zongcheng Ling, who studies the formation and evolution of planetary bodies at Shandong University in Weihai, and is a member of the mission's science team. "We scientists are very happy" to have the chance to visit the far side, says Ling. The crater is part of the South Pole–Aitken basin, the largest known impact structure in the Solar System and the oldest on the Moon.

"It is a key area to answer several important questions about the early history of the Moon, including its internal structure and thermal evolution," says Bo Wu, a geoinformatician at Hong Kong Polytechnic University, who helped to describe the topography and geomorphology of this site.

The Chang'e-4 rover will map the region surrounding the landing site. It will also measure the thickness and shape of the subsurface layers using ground-penetrating radar, and measure the mineral composition at the

"It is a key area to answer important questions about the early history of the Moon."

surface with a near-red and infrared spectrometer, which could help geologists to understand the processes involved in the Moon's early evolution.

Because the far side of the Moon never faces Earth, CNSA mission control won't be able to communicate directly with the craft once it has landed. In May, China launched a communications satellite called Queqiao to orbit beyond the Moon and act as a relay station.

The Chang'e-4 rover and lander will carry out some unique experiments. One of those will test whether potato and thale cress (*Arabidopsis*) seeds sprout and photosynthesize in a sealed, climate-controlled environment in the low gravity on the lunar surface.

"When we take the step towards long-term human habitation on the Moon or Mars, we will need greenhouse facilities to support us, and will need to live in something like a biosphere," says Anna-Lisa Paul, a horticultural scientist at the University of Florida in Gainesville.

The proposed test will seek to verify previous studies carried out on the International Space Station, says John Kiss, a space biologist at the University of North Carolina Greensboro. These found that potatoes and thale cress can grow normally in controlled ecosystems in lower gravity than that on Earth, but not in gravity as low as that on the Moon.

The lander's radio-astronomy experiments will target parts of the Milky Way that are poorly understood, such as the gases between stars, and the magnetic fields that propagate after a star's death. A radio spectrometer, built by the Chinese Academy of Sciences, will collect electromagnetic data between 0.1 and 40 megahertz to create a map of low-frequency radiation from the night sky.

Capturing these measurements from Earth is difficult because Earth's atmosphere mostly blocks such radiation, says Heino Falcke, a radio astronomer at Radboud University Nijmegen in the Netherlands, and a member of the Dutch team that has built a spectrometer carried on the Queqiao satellite. "We have completely blurred vision at low frequencies," he says. Astronomers will use these data to better understand how energy released by dying stars heats up the gases between them, which could affect how stars form, says Falcke.

They are also interested in this spectrum of radiation in order to study the first few hundred million years of the Universe, a time before the formation of galaxies and stars. The data could help the researchers filter out background noise that could be hiding a signal from this time period. If found, that signal could reveal information about the distribution of ordinary matter compared with dark matter in the Universe. But even with the help of the Moon lander, it is not certain that these experiments will detect the signal, says Falcke. "It is a first step."

China's next venture to the Moon will be even more ambitious. Chang'e-5, scheduled to launch in 2019, will try to bring samples from the Moon back to Earth. ■

ENVIRONMENT

Venice's massive flood gates could wreck ecosystem

Proposed system comes under renewed scrutiny following recent floods.

BY LOU DEL BELLO

An ambitious plan to prevent the Italian city of Venice from being swallowed by the sea could spell disaster for the lagoon that surrounds it.

The system, called MOSE (from the Italian for Experimental Electromechanical Module), is in the final stages of construction and is expected to be completed in 2022. It would consist of a complex network of 78 flap gates designed to separate the lagoon that hosts the city from the Adriatic Sea in times of high tide that would otherwise lead to flooding.

But according to modelling studies, as sea levels keep rising, MOSE will become less effective at preventing flooding in the city while increasingly compromising the lagoon's delicate ecosystem.

The environmental impacts of the €6-billion (US\$6.5-billion) project have been a sore spot since its conception in 1992. MOSE has come under renewed scrutiny in recent weeks, following both exceptionally extensive flooding this October that submerged large parts of the city in 156 centimetres of water, and the release of new data and simulations highlighting the city's vulnerability to the rising sea (L. Reimann *et al.* *Nature Commun.* **9**, 4161; 2018).

Researchers now say that MOSE's

monumental structure, whose gates lift to provide an artificial barrier to the sea and stem unusually high tides, will damage the lagoon's ecosystem and the maritime economy in just a few decades.

Luigi D'Alpaos, a hydrologist at the University of Padua in Italy, says that the problem is not the structure itself, but how often the gates would need to be closed when the sea level rises and exceptionally high tides become more frequent.

D'Alpaos simulated the potential outcomes of different sea levels by looking at all high tides between 2000 and 2012. Earlier this year, his team found that with a sea-level rise of 50 centimetres — the level predicted by the latest Intergovernmental Panel on Climate Change report (see go.nature.com/2rkasia) — the lagoon would have to be closed for up to 187 days each year, occasionally for weeks at a time (R. Mel and L. D'Alpaos *Atti dell'Istituto Veneto di Scienze, Lettere ed Arti* **176**, 1–58; 2018). This would quickly deplete the lagoon's oxygen, they say, and in turn harm the populations of fish and many bird species nesting in the area, such as flamingo, peregrine falcon, black swan and cattle egret. This pits MOSE's anti-flood measures against conservation efforts.

"To save the lagoon, we would have to open the gates — removing the only barrier against

flooding," says D'Alpaos.

To avoid oxygen depletion in the lagoon, the Venezia Nuova, a consortium tasked with the implementation of the project, says that MOSE will be activated only on days when the water level rises 110 centimetres above average.

But this measure is unlikely to spare the city from regular flooding, say scientists, including D'Alpaos. Floods caused by water levels between 70 and 100 centimetres above average are common, and inundate the city's iconic St Mark's Square and other attractions for long periods.

The floods this October lasted for 30 hours. Had MOSE been active, the gates would have been raised for 20 hours of that period, says Monica Ambrosini, a spokesperson for the Venezia Nuova. Models show that future floods are expected to occur more frequently and to last for days at a time, which would require longer closures.

THINKING DEEPER

Andreina Zitelli, an environmental scientist at the University of Venice, who has criticized MOSE's green credentials, is one of several people who have been exploring alternatives.

One such proposal, originally dating back to the 1970s, involves injecting fluid cement, or even water, beneath the city to raise it above floodwater thresholds. Officials tested this technique on the small island of Poveglia in the Venetian lagoon in the 1970s. When workers injected a cement compound 10 metres underground, the tactic raised the island by 10 centimetres.

Other flood-adaptation proposals include injecting water hundreds of metres underground, through 12 wells surrounding Venice, mirroring a method widely used to stabilize oil rigs as they extract fluid.

The science behind this idea is solid and widely tested by oil companies worldwide, says Georg Umgiesser, an oceanographer with the Italian National Research Council, in Venice.

"The case of Venice would be more complex, because the city has a fragile structure and has already experienced 25 centimetres of subsidence, so any intervention should correct that problem first," says Umgiesser. He adds that too much money and time have been invested in MOSE to abandon the project now, "but once it's completed, at that point we can think about something else". ■



Severe floods hit Venice in October; in future, they will occur more frequently and last for longer.

STEFANO MAZZOLA/ANWAKENING/GETTY



Greater Tehran is home to about 13 million people.

IRAN

Tehran's drastic sinking exposed

Satellites reveal city is subsiding by 25 centimetres a year.

BY KATE RAVILIOUS

Tehran, western Asia's most populous city, is sinking.

Now, detailed satellite images reveal the extent of the problem, showing that some parts of the Iranian capital are falling by as much as 25 centimetres a year, and that the collapse is spreading to encompass the city's international airport (see 'Feeling low').

Geoscientists Mahdi Motagh and Mahmud Haghshenas Haghighi, both at the GFZ German Research Centre for Geosciences in Potsdam, used satellite data to monitor subsidence¹ across the Tehran region between 2003 and 2017.

Previous work² had shown that the capital is sinking, and had linked the subsidence to the depletion of groundwater aquifers, which are being sucked dry to irrigate nearby farmland and serve Greater Tehran's 13 million or so residents³.

The latest data put new figures on the problem. The western Tehran Plain — a mix of Tehran's urban sprawl, satellite cities and agricultural land — is subsiding at a rate of 25 centimetres per year, and the Varamin Plain, an agricultural region to the southeast of the city, is falling at a similar rate. The city's international airport — located southwest of the centre — is sinking by 5 centimetres annually.

"These are amongst some of the highest

current rates of subsidence in the world," says Roberto Tomás, an engineer at the University of Alicante in Spain.

LOSING GROUND

Subsidence, caused by growing populations and increased extraction of underground water, oil and gas, is a problem in cities globally. Previous satellite measurements have shown, for example, that some areas of Jakarta are sinking more than 20 centimetres per

year⁴, and the San Joaquin Valley in California — home to several cities — by up to 60 centimetres per year.

The latest study¹, which has been accepted for publication in *Remote Sensing of Environment*, estimates that around 10% of Tehran's urban area is affected, along with many satellite towns and villages to the city's southwest. "When walking around these areas, we see uneven street surfaces, shifted curbs, cracks in the walls and even tilted buildings, some of which have had to be demolished," says Motagh.

Huge fissures — several kilometres long and up to 4 metres wide and deep — have opened up in the land to the southeast of Tehran, and some are threatening to topple power-transmission lines and buckle railways.

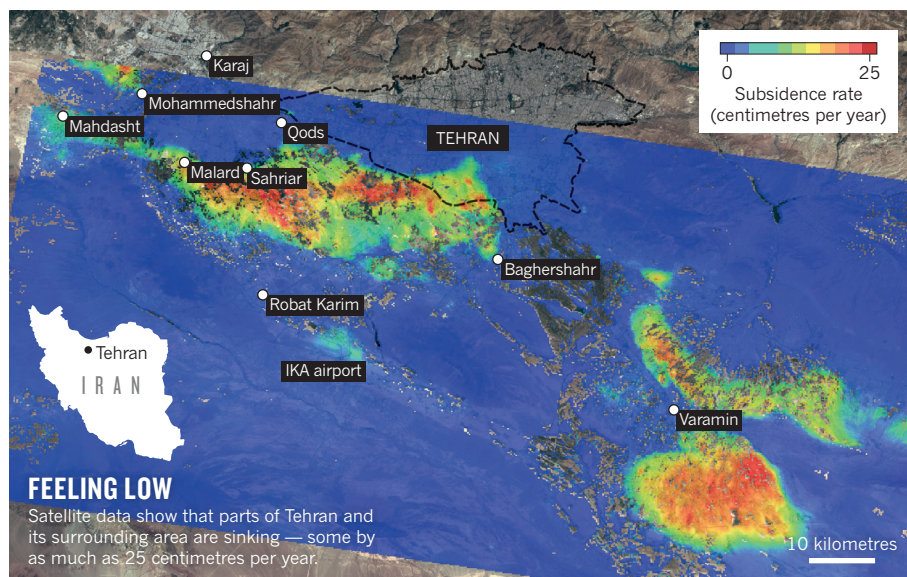
And the growth of underground cracks sometimes produces sudden sinkholes. "One farmer I met was locked up for hours when the ground gave way beneath him and he fell into a 6-metre-deep crack," says Ali Beitollahi, head of engineering seismology at the Building and Housing Research Center in Tehran. Some farmland is becoming unviable, because the cracks drain irrigation water from the surface and leave crops parched.

DRY EARTH IN IRAN

Surveys carried out over the past year by Beitollahi and his colleagues estimate that the areas with significant subsidence in and around Tehran host 120 kilometres of railway, 2,300 kilometres of road, 21 bridges, 30 kilometres of oil pipeline, 200 kilometres of gas pipeline, 70 kilometres of high-voltage electricity lines and more than 250,000 buildings.

Motagh and Haghshenas Haghighi's data show how the subsidence has marched steadily eastwards since 2003, starting with agricultural land and encroaching on the urban fringes of the city. Another subsidence zone is creeping towards Tehran's airport.

A combination of population growth ►



MEHMETO/ALAMY

SOURCE: M. HAGHSHENAS HAGHIGHI & M. MOTAGH (ANALYSIS)/COPERNICUS SENTINEL 2015–17 (DATA)

► — the city's population has doubled in the past 40 years — droughts and large dams, which capture rainwater and prevent aquifers from recharging, has exacerbated the problem.

The authorities are fighting a losing battle as they try to regulate water extraction. Beitollahi thinks that some 100,000 illegal wells have been blocked across Iran, but that an estimated 30,000 are still in operation

across Greater Tehran.

The sinking that has already happened might be irreversible, the study hints. By looking at water-depth measurements from wells in the affected areas, the researchers found that the ground is failing to bounce back, even after rainfall, which suggests that the porosity of the rock has been permanently lost. That loss could lead to more flash flooding, says

Linlin Ge, an engineer at the University of New South Wales in Sydney, Australia, because without pores in the rock, the water no longer has anywhere to go. ■

1. Haghshenas Haghighi, M. & Motagh, M. *Remote Sens. Environ.* (in the press).
2. Pirouzi, A. & Eslami, A. *Int. J. Geo-Eng.* **8**, 30 (2017).
3. Motagh, M. *et al. Geophys. Res.* **35**, L16403 (2008).
4. Abidin, H. Z. *et al. Nat. Hazards* **59**, 1753 (2011).

ARCHAEOLOGY

Geologists track ancient sites' bullet wounds

The ultimate goal is to inform efforts to conserve or repair heritage sites.

BY SARAH WILD

In 2015, Lisa Mol stared at a series of satellite images, distraught. The before-and-after pictures showed how the Islamist terrorist group ISIS had damaged the ancient Syrian city of Palmyra with explosives and bulldozers. An oasis in the desert, Palmyra had been a cultural meeting place in the first and second centuries AD, and contained the fingerprints of many civilizations.

"Seeing that deliberate destruction pushed me into taking action," says Mol, a geomorphologist at the University of the West of England in Bristol. "I am not a lawyer, I cannot do anything medical, but I do know rocks."

Mol, who specializes in rock art and rock deterioration, is now spearheading an initiative — the first of its kind — to quantify and catalogue the impacts of bullets in rock at a heritage site in the Middle East. The eventual goal is to inform efforts that aim to conserve or repair such sites.

Typically, people look at the effects of conflict on a site in its totality, rather than at individual instances of damage, says Robert Bewley, who specializes in endangered archaeology at the University of Oxford, UK. "The science into what's going on is very important," he says. "If there is no science, people may say, 'Let's just slap concrete over it and it will be fine.' It won't."

BALLISTIC EXPEDITION

Satellite imagery has been used to identify damage in conflict areas, such as Syria and Libya. But there is a dearth of information about how stone structures weather after ballistic damage, despite the fact that ancient sites are often casualties of war — and have been for centuries. "I saw something that needed doing, and built up a team," Mol says.

Mol's team, comprising a palaeontologist,



Bullet damage on rock art at Wadi Rum, a site of prehistoric human settlement in Jordan.

two geomorphologists, a heritage specialist and an archaeologist, returned in September from an expedition to Wadi Rum, a heritage site in Jordan. Wadi Rum is home to rock paintings, engravings and archaeological remains that document millennia of human habitation, and it wears the scars of conflicts old and new. The rocks' physical characteristics, or lithology, are also similar to those in areas such as Syria, where safety issues are too great for researchers to make expeditions.

The team hopes ultimately to develop step-by-step guidelines for locals to identify and catalogue ballistic damage to heritage sites — for use in Jordan and beyond. Residents could record and communicate their findings using an information sheet, or send images to researchers

by e-mail or through an app, says Mol.

But the researchers must first determine which stone properties are most crucial for tracking ballistic damage and environmental degradation. "We can't simplify to that level without the high-level scientific understanding," says Mol.

The bullet damage at Wadi Rum spans decades, from guerilla conflict in the early twentieth century to damage from AK-47 guns in the past few months, thought to have been caused by people using rocks for target practice. Over the decades, munitions have changed — as has the extent of the harm they cause. How badly weathering worsens after ballistic hits depends on many factors, including weapon type, rock composition and climate. This degradation

LUCY CLARKE

can be as harmful as the initial bullets, says Bewley, but is not well studied.

During their expedition, Mol's team collected data on the surface hardness, resistivity and permeability of rocks, both at points of impact and in undamaged rock. They will combine these data with 3D images of the surface morphology to calculate the size, depth and shape of impacts, as well as the fractures that run along the surface.

In Mol's lab, researchers will shoot guns at rocks to test the microstresses caused by bullet impacts from different weapons — and compare the results with the data from Wadi Rum to work out which weapons created which impacts, and how damage plays out in the rock.

But even with rich on-the-ground data, it can be difficult to determine exactly who shot at the sites and when. Historical conflict is a likely culprit for some of the damage at Wadi Rum, says Kaelin Groom, a geographer at Arizona State University in Tempe who is on the team. But many of the impacts are known to be from acts of vandalism. The researchers also interviewed local residents to narrow the time frames and identify possible shooters.

Heinz Ruther, a geoinformatician at the University of Cape Town in South Africa, says that he's not aware of other researchers doing such ballistic work. Being able to quantify the extent of conflict damage to heritage sites

would be very relevant, says Ruther. So many buildings are affected or destroyed by war, he says, but partial damage is seldom considered.

MEMORIAL SCARS

But there is more to heritage conservation than scientific understanding. Local stories and knowledge about the bullet impacts affect how the sites should be conserved, says Rachel

“If there is no science, people may say, ‘Let’s just slap concrete over it and it will be fine.’”

King, an archaeologist at University College London who was part of the expedition. Some residents think that certain bullet damage should not be repaired but should instead stand as a warning against vandalism or as a reminder of the conflict that caused it.

Mohammad Dmayan Al-Zalabiah's family has lived in Wadi Rum since the early nineteenth century. A tour guide, Al-Zalabiah was part of a US programme aimed at managing cultural heritage resources in Jordan. He worked to create a database of local rock art and inscriptions at the site, and helped Mol and her team to collect data about bullet damage. He thinks that the ballistic research has value for the community, because it highlights the extent of bullet damage and dissuades vandals.

“You can't understand something as complex as the physical damage to heritage,” says Groom, “without social outreach, ethnography and geology.” ■

CLARIFICATION

Some phrasing in the News Feature ‘Does science have a bullying problem?’ (*Nature* **563**, 616–618; 2018) did not make it clear that Nazneen Rahman resigned from the Institute of Cancer Research before the Wellcome Trust revoked her funding.

CORRECTIONS

The News Feature ‘Against all odds: science in the Palestinian territories’ (*Nature* **563**, 308–311; 2018) located Mohammad Herzallah at the wrong campus. He is in Newark, not Piscataway.

The News story ‘Mystery supernova known as ‘Cow’ spills its secrets’ (*Nature* **563**, 168–169; 2018) omitted to cite an important reference for the discovery: S. J. Prentice *et al.* *Astrophys. J.* **865**, L3 (2018).

The News Feature ‘The sun dimmers’ (*Nature* **563**, 613–615; 2018) said that David Keith received money from the Bill & Melinda Gates Foundation. In fact, the money came directly from Bill Gates.

npj | Clean Water

Call for Papers

Publishing cutting-edge research on water treatment

npj Clean Water is an online-only, open access journal, dedicated to publishing significant research that ensures the supply of clean water to populations worldwide.

Publishing with npj Clean Water offers authors a number of benefits, including:

- High visibility and wide dissemination
- Global reach and discoverability via nature.com
- Compliance with open access funding mandates
- Comprehensive and rigorous peer review by experts in your field

Published in partnership with



EDITOR-IN-CHIEF

Professor Eric M.V. Hoek, Ph.D.

Chief Executive Officer and Founder at Water Planet, Inc.

Department of Civil & Environmental Engineering, University of California, Los Angeles, USA

Part of the Nature Partner Journals series

npj | nature partner journals

nature.com/npjcleanwater

nature research

PICTURES OF WORLDS TO COME

Images of newborn planets, still swaddled in gas and dust, are challenging theories about how worlds take shape.

BY REBECCA BOYLE

Some 100,000 years ago, when Neanderthals still occupied the caves of southern Europe, a star was born. It appeared when a ball of gas collapsed and ignited within a stellar factory known as the Taurus Molecular Cloud. Then, leftover material began to cool and coalesce around it, forming dust grains and a hazy envelope of gas.

In September 2014, some of the light from that hot young star and its surroundings landed inside 66 silvery parabolas perched on a plateau in Chile's Atacama desert — the driest on Earth. The photons had taken 450 years to make the journey. Astronomers were waiting. They were conducting a test of the Atacama Large Millimeter/submillimeter Array (ALMA), which features radio antennas separated by distances of up to 15 kilometres. With such long spans between them, the antennas work as a high-resolution receiver that can discern cool objects less than a millimetre across.

When the telescope team trained ALMA on the young star, named HL Tauri, they expected to see a bright smear of dust and gas. Instead, when ALMA's supercomputer stitched together those photons, the image resolved into a disk with a well-defined ring structure, with gaps seemingly etched by small, infant planets orbiting a central star. It looked like a furry, orange Saturn¹. It looked like nothing astronomers had ever seen.

"I kept flipping through their paper, and I was like, 'Where is the real image? This is obviously a model,'" says Kate Follette, an astronomer at Amherst College in Massachusetts.

What the researchers had captured was a picture of a planetary nursery — where baby planets were forming in a disk of gas and dust around HL Tauri. This observation marked the start of a revolution in the burgeoning field of planetary-disk imaging. In the four years since, astronomers have captured 'baby pictures' of numerous other systems. These planet-forming regions exhibit a wide variety of patterns. Some are neat ovals, with lanes as clearly defined as those of a race track. Others look like

galaxies in miniature, with swirling arms that branch off into open arcs.

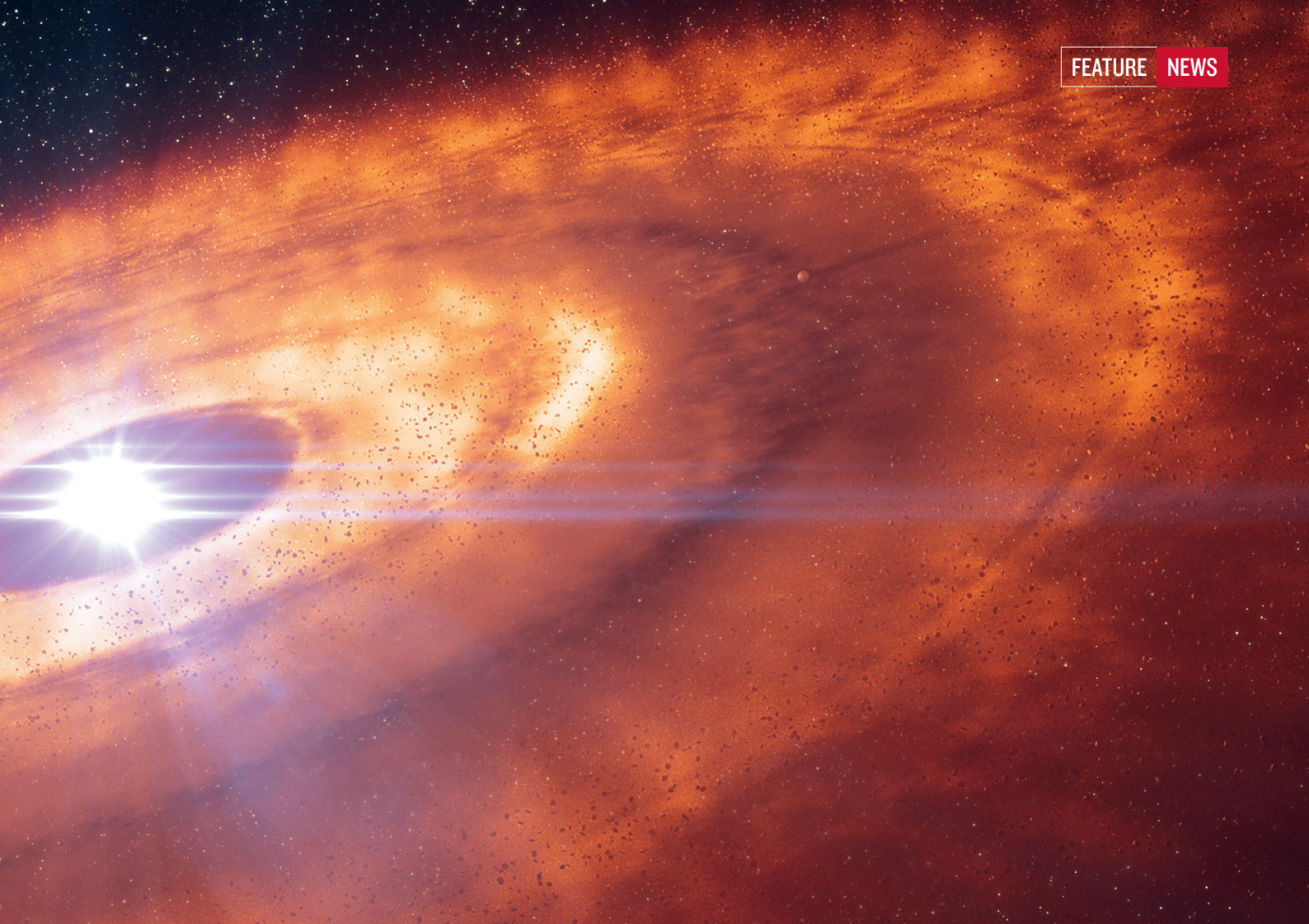
The latest observations, including results announced in April and July, have revealed planets in the process of being sculpted, with dust and gas flowing onto bulbous, red-hot infant worlds^{2,3}.

But as the menagerie of young planetary systems grows, researchers are struggling to square their observations with current theories on how our Solar System and others formed. Such ideas have been in turmoil ever since astronomers started discovering planets around distant stars — a list that now numbers in the thousands. The Solar System has rocky planets near the Sun and giant gas balls farther out, but the panoply of exoplanets obeys no tidy patterns. And the rule book for world-building is getting more complicated as researchers find evidence of planets in the process of being born. Still, astronomers hope that witnessing such birth pangs will shed light on how all planetary systems, including our own, came to be. "We see all kinds of structure in these disks, even at very young ages," says Follette. "Even younger than we classically thought planets should form."

COLLISIONS AND CURDLING

The prevailing theory of how the Solar System formed goes back to the German philosopher Immanuel Kant. In 1755, he imagined the Sun and planets arising from a nebulous cloud of gas and dust that slowly collapsed and flattened. Today, the widely accepted general model for how the process unfolded holds that the Sun collapsed inside a molecular cloud, a star factory full of gas molecules. A ring of gas and dust would have remained after the star formed, cooling and progressively condensing into bigger grains, then into larger, asteroid-sized bodies called planetesimals, and ultimately into planets.

Theorists have been refining the particulars of the process since the 1970s, taking into account the distribution of planets in the Solar System and the chemical components of meteorites — crumbs from the Solar



ESO/L. CALÇADA

System's formation. By the early 2000s, they had settled on two distinct scenarios for making rocky planets and gas giants (see 'Attractive scenarios').

In one theory, called core accretion, rocky material violently smacks together, melts, coagulates and forms larger bodies, gradually creating protoplanets — compact embryonic worlds several thousand kilometres across. With their gravitational heft, protoplanets can attract a huge envelope of gas as they orbit through the planetary disk. This could enable them to metamorphose into the core of a giant planet, such as Jupiter; alternatively, their growth might ultimately stall at the rock-ball stage, as happened with Earth, Mars and the other terrestrial planets.

Others theorized that the Solar System was forged not through violent collisions, but instead by a kind of curdling. In this scenario, called the streaming instability, gas and dust surrounding a star cool off quickly and begin drifting, becoming concentrated and collapsing under their own gravity. The centimetre-scale dust and ice in the disk forms agglomerations that grow into larger, denser bodies between 1 and 100 kilometres across. Then, through other processes, these grow into larger planetary embryos and, eventually, planets.

But neither of these ideas can quite explain

An artist's impression of a planetary nursery, in which growing planets etch rings in the disk of dust and gas around a young star.

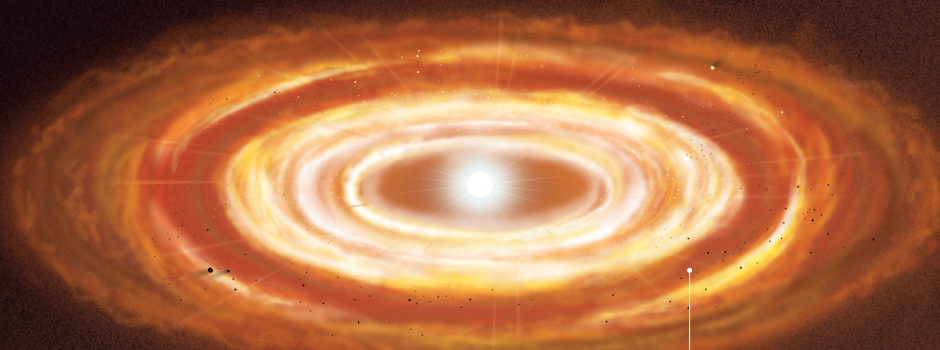
the Universe we see. Take Jupiter, which contains the vast majority of the material left behind from the Sun's birth. Among the biggest questions is how the planet could have quickly grown a core big enough to hoover up the bulk of its mass; collisions between planetesimals would take many millions of years. But theorists reckon that the 'natal disk' of dust and gas that surrounded the young Sun would have disappeared 1 million to 10 million years after it formed, as gas dissipated and dust spiralled onto the star. (Compounding the problem, NASA's Juno probe recently revealed that Jupiter's core is even bigger than expected, meaning that the formation process must have been extremely fast.) Jupiter's location is also hard to explain. Theorists have speculated since the 1970s that planets might migrate from one orbit to another as they form or jostle with other burgeoning planets.

The cracks in planet-formation theories only got worse in the mid-2000s, as discoveries of other planetary systems began rolling in. Some stars have large planets that complete their orbits in just a few days. Other planets circle their hosts at distances that make Jupiter seem like the Sun's next-door neighbour. Although simulations are growing more complex as hardware and software improve, neither core-accretion nor streaming-instability models do a good job of explaining how such huge worlds are formed, and at such disparate distances from their stars.

One scenario that could account for far-out planets emerged in 2012. Astronomers Anders Johansen and Michiel Lambrechts at Lund University, Sweden, devised a variation on the core-accretion and streaming-instability scenarios. In their theory, dubbed pebble accretion, leftover star-forming material assembles as loose collections of dust and pebbles. Already-formed planetesimals swim among them, and then grow quickly by accumulating more pebbles, much as a snowball gets bigger as it rolls downhill. In this scenario, Johansen says, a planet would start out at the edges of a star's natal disk and gather up pebbles as it migrates inwards. Depending on gravitational interactions between worlds, it

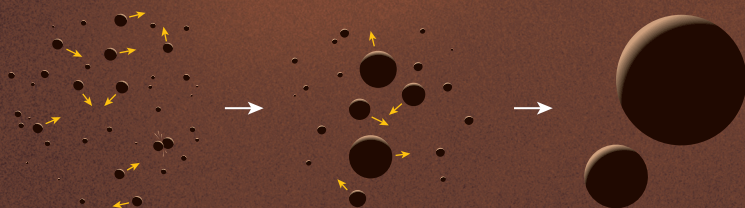
ATTRACTIVE SCENARIOS

Various theories have been proposed to explain how planets come to be. Many focus on the crucial period right after the birth of a star, when the dust and gas surrounding it somehow transform from a relatively uniform disk into planetary embryos called protoplanets — objects several thousand kilometres wide that ultimately form the cores of giant gas planets and the bulk of smaller, rocky ones.



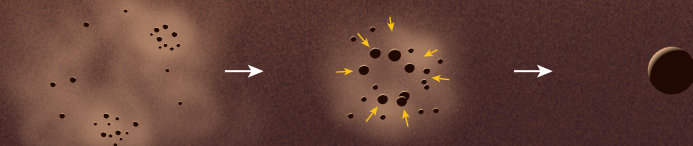
CORE ACCRETION

In this early theory, protoplanets form through a series of violent strikes, as bits of dust and then progressively larger objects are gravitationally attracted to one another and collide — and, in many cases, merge.



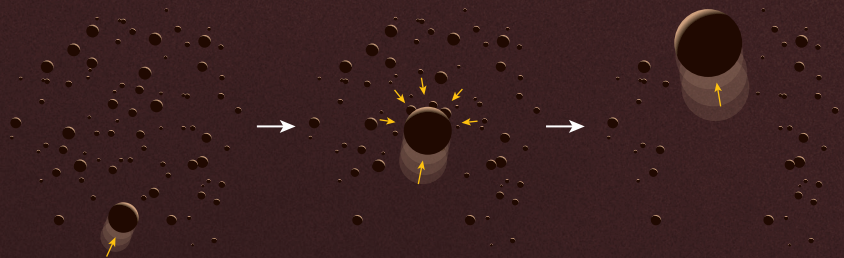
STREAMING INSTABILITY

Another scenario proposes instead a gravitational collapse, in which density variations cause solid lumps to collapse into asteroid-sized 'planetesimals', which then grow into protoplanets by other means.



PEBBLE ACCRETION

One way in which such planetesimals might grow is described in a third theory, which proposes that larger pieces of debris draw in smaller 'pebbles' as they move through the disk. Thanks to both gravitational and hydrodynamic interactions, these small pieces adhere like snow to a rolling snowball.



could end up either very close to its host star, or far removed from it. Astronomers think that Jupiter and Saturn might have undergone such a migration early in the life of the Solar System.

Pebble accretion has quickly gained popularity as a way of explaining systems such as HL Tauri, whose dark rings, etched in luminous dust, seem to harbour planets less than 100,000 years old. "These dark rings probably have young planets" in them, says Matthew Clement, an astronomer at the University of Oklahoma in Norman. "This has been really inspirational for us. It's confirmation, in a way, that planets grow really fast."

TALLYING IT UP

Although pebble accretion could explain how planets get big fast, it doesn't provide as much insight into how the seed of a planet — the start of the snowball — forms in the first place.

The challenge is bridging the gap between centimetre-scale bits of dust and Moon-sized objects. Older simulations assumed that dust and gas moved together. "When people did this problem historically, they always assumed the dust and gas were perfectly locked to each other," says Philip Hopkins, an astronomer at the California Institute of Technology in Pasadena.

He and Jono Squire, a postdoctoral researcher in his lab, have been revising models to separate the two, exploring complex interactions in a protoplanetary disk that can cause gas to swirl around dust grains in the same way as water eddies around sticks floating in a stream⁴. These redirected gas flows quickly become turbulent and unstable, forcing dust to clump together like flood debris. Such modelling could help to shed light on the fundamentals of planetesimal clumping, Hopkins says. "This could really change the story."

But as theorists tinker with accreting pebbles and swirling gas, another problem is lurking in the background. In 2013, astrophysicist Subhanjoy Mohanty of Imperial College London and astronomer Jane Greaves, now at Cardiff University, UK, published an initial survey of protoplanetary disks in the Taurus Molecular Cloud⁵. The observatories they used were not powerful enough to clearly resolve grooves in disks like those that ALMA saw around HL Tauri, but when the researchers tallied up how much gas and dust seemed to be present, they found that intermediate-sized stars had disks that packed much less mass than expected.

This summer, astronomer Carlo Manara at the European Southern Observatory (ESO) in Garching, Germany, took another look, and found this to be true throughout the Milky Way⁶. Protoplanetary disks have just a fraction — sometimes as little as 1% — of the combined mass of exoplanets orbiting similar stars, he found. This would mean that planetary systems are bigger than the stuff used to make them.

Whatever the explanation for this seemingly impossible scenario, theorists will have to

FROM TOP: ALMA (ESO/NAOJ/NRAO); ESO/H. AVENHAUS ET AL./DARTT-S
COLLABORATION; ESO, T. STOLKER ET AL.; ESO/A. MÜLLER ET AL.

grapple with the implications. To account for exoplanet observations, they have generally started with vast quantities of material. “You need a huge amount of mass in the disk [for it] to exert gravity on itself to act like a seed, and collapse on itself,” Greaves says.

It is possible that there is more here than meets the eye. There could, for example, be material in the disk that is difficult for telescopes to catch. Or, as Manara and his colleague Alessandro Morbidelli, a dynamicist at the Côte d’Azur Observatory in Nice, France, suggest, astronomers might be seeing only a snapshot; stars might be accreting new material from outside the protoplanetary disk, from the molecular clouds that forged them.

This theft could be hard to spot. But in research published in 2017, astrophysicist Hsi-Wei Yen at the ESO and his colleagues described two gas streams that seem to be connected to HL Tauri’s disk — although they couldn’t tell whether the gas was flowing towards or away from the star⁷. If it were heading towards the star, Morbidelli says, the inflowing gas would have wide impacts, because it would also affect factors such as the disk’s temperature, density and magnetism. Finding evidence of such flows suggests that stars and planets are not isolated from the larger cosmos as they form and grow. “The disk is not in a box,” he says, “and this is also a revolution in our thinking about disks.”

PLANETARY MENAGERIE

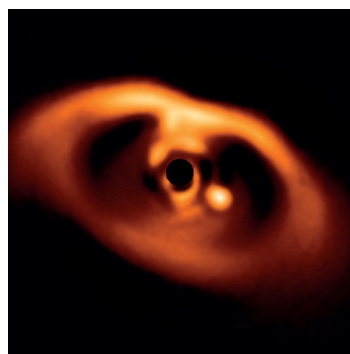
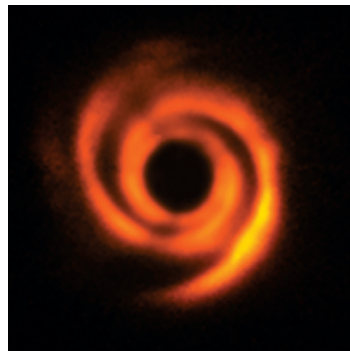
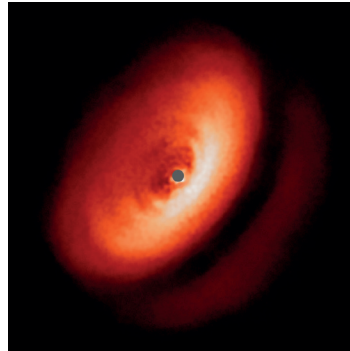
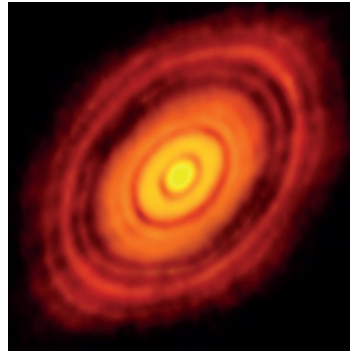
As if theorists did not already have enough to grapple with, observations of planetary nurseries continue to pile up. The latest findings lend weight to the idea that planets are forming early in the lives of their stars, and at distances from them that vary widely.

And it’s not just ALMA that’s been supplying images. Astronomers have also turned to the SPHERE instrument mounted on the ESO’s Very Large Telescope. This, too, is in the Atacama desert, about a six-hour drive south of ALMA. SPHERE has a system that can cancel out the blurring effects of the atmosphere and a filter that blocks starlight. In April, astronomers announced that they had used it to capture a diverse array of disks around eight young Sun-like stars². Some resembled wide platters, some had distinct racetrack-like ovals, and one resembled a galaxy with jets streaming from its centre. Such diversity suggests that planet-forming is a complex process yielding many possible outcomes.

Just two months later, news came that ALMA had been used to snap what might be the youngest exoplanets ever seen, orbiting a 4-million-year-old star about 100 parsecs (330 light years) from Earth^{8,9}. ALMA, which is at its most sensitive when viewing small, cool objects, cannot see starlight reflecting off the planets directly. But the swirl of carbon monoxide gas in the disk of the star suggests that three planets — each roughly the mass of Jupiter — are in orbit, forcing gas to flow around them, as rocks control the flow of a stream.

Not to be outdone, astronomers who had turned SPHERE towards another young star, called PDS 70, managed to nab a direct image of a gas giant. The planet orbits its star about four times farther than Jupiter does from the Sun, and is still gobbling up material from its natal disk of dust and gas³. The observation confirms the prediction that gas planets such as Jupiter form at vast separations from their stars.

Another instrument, the Gemini Planet Imager (GPI), which is mounted on the Gemini South Telescope in Chile’s Andean foothills, has also been capturing disks with planets embedded in them, including a large gas giant that seems to support the core-accretion scenario of planet formation¹⁰. As more observations roll in, lingering



From top, an ALMA image of gaps etched by growing planets in a disk of material surrounding the young star HL Tauri; SPHERE captures a dusty disk around IM Lupi; a spiralling disk around HD 135344B; the glow of a gas giant around PDS 70.

doubts about whether these young nurseries are really cradling planets — and not, say, displaying instabilities in their disks — are being put to rest. “Almost all of the features that we see can be explained most easily by planets,” says Follette, who works on the GPI.

But the latest findings are also showing astronomers that the Universe is much more complex and richly detailed than even our most advanced theories can predict. Several astronomers are realizing that the theoretical work they were doing a decade ago is no longer valid, but they are still not sure how to fix it.

“There’s always that aspect; I’m sad that the stuff I did in the past isn’t right any more. But the truth is, it was never right,” says Sean Raymond, an astronomer at the Bordeaux Astrophysics Laboratory in France. “It was hopefully a step forward.”

Observations might be of limited use in resolving the picture. ALMA and other radio observatories can see the dust and gas surrounding young stars, and optical instruments such as SPHERE and the GPI can see the disks and planets embedded in them, lit up with reflected starlight. But the range between tiny debris and 1,000-kilometre worlds will remain invisible.

Still, current and future telescopes could help to fill in some gaps. Astronomers could reach beyond ALMA’s millimetre-scale vision to the centimetre range, Greaves says, with higher-resolution radio observations from telescopes such as the United Kingdom’s Merlin array — as well as from the forthcoming Square Kilometre Array, due to be hosted in South Africa and western Australia. Such observations could partly bridge the span between dust and protoplanet. Greaves eagerly anticipates the possibility of finding centimetre-scale material swirling around what could be future rocky planets. “Seeing a spot in a disk that indicated an Earth forming at an Earth-like distance from its star — that’s the new holy grail, at least for me.”

With the observation of protoplanetary disks still in its infancy, the full story of planet-making will probably be more complicated than anyone expects, and ideas could well be overturned and then overturned again. “Case in point, it looks like the Solar System isn’t even the most common-looking system out there. We’re a little weird,” says Clement. “It turns out there is a lot of complexity out there.” ■

Rebecca Boyle is a freelance science journalist based in St Louis, Missouri.

1. ALMA Partnership *Astrophys. J. Lett.* **808**, L3 (2015).
2. Avenhaus, H. et al. *Astrophys. J.* **863**, 44 (2018).
3. Keppler, M. et al. *Astron. Astrophys.* **617**, A44 (2018).
4. Hopkins, P. & Squire, J. *Mon. Not. R. Astron. Soc.* **479**, 4681–4719 (2018).
5. Mohanty, S. et al. *Astrophys. J.* **773**, 168 (2013).
6. Manara, C. F., Morbidelli, A. & Guillot, T. *Astron. Astrophys.* **618**, L3 (2018).
7. Yen, H.-W. et al. *Astron. Astrophys.* **608**, A134 (2017).
8. Pinte, C. et al. *Astrophys. J.* **860**, L13 (2018).
9. Teague, R. et al. *Astrophys. J.* **860**, L12 (2018).
10. Macintosh, B. et al. *Science* **350**, 64–67 (2015).



AFRICA'S SILENT EPIDEMIC

Hepatitis now kills more people worldwide than HIV, tuberculosis or malaria. Tackling the hepatitis B virus in Africa is key to fighting back.

BY IAN GRABER-STIEHL

Nuru was prepared for the worst when she went to get screened for HIV eight years ago. After caring for her mother in Uganda, who died as a result of the virus, Nuru moved to the United Kingdom to study, and decided to take her health into her own hands. “I was ready to be told I had HIV,” she says. “I felt, ‘That’s okay. I’ve looked up to my mother.’”

SVEN TORFINN/PANOS

What she didn’t expect was to be diagnosed with a different viral infection altogether: hepatitis B. “The way the health worker delivered it to me, it was like, ‘It’s worse than HIV’. I was confused, I was suicidal,” says Nuru (who asked that her real name not be used for this article). “I just didn’t understand what it was because no one ever talks about hep B — they talk about HIV. That’s well researched, it’s well talked about, well documented. It’s all over the television. But hep B is not.”

The hepatitis B virus (HBV), which spreads through blood and bodily fluids and invades liver cells, is thought to kill just under 1 million people every year around the world, mostly from cancer or scarring (cirrhosis) of the liver. HBV is less likely to be fatal than HIV, and many people who carry the virus don’t have symptoms. But because more than 250 million people live with chronic HBV infections, more than 7 times the number with HIV, its global death toll now rivals that of the more-feared virus.

Hepatitis — or liver inflammation — is caused by a number of viruses, but types B and C are associated with the most deaths. In 2016, the most recent year for which estimates are available, the number of deaths worldwide from viral hepatitis rose to 1.4 million, outstripping those from tuberculosis, HIV or malaria individually (see “The burden of hepatitis B”).

This is despite the fact that HBV infection can be prevented by vaccination early in childhood and treated with the same antiretroviral drugs used to combat HIV. “HIV has been an acute pandemic with resources thrown at it. That’s a completely different picture than hep B, which has travelled with humankind for tens of thousands of years — and by dint of that invisible carriage, has never had that injection of political advocacy, funding, energy and education that’s gone into HIV,” says Philippa Matthews, an immunologist at the University of Oxford,

A market in Uganda — a nation where 6% of people carry hepatitis B. UK, who studies viral infections such as HBV. Researchers and health workers are now hoping to change that. Two years ago, the World Health Assembly endorsed a World Health Organization (WHO) strategy to eliminate hepatitis as a public-health threat by 2030, which the WHO defined as reducing new infections by 90% and deaths by 65%.

A major focus is to combat the growing HBV crisis in sub-Saharan Africa. Other high-risk regions, such as the Western Pacific (which stretches from China to New Zealand), have long inoculated children against the virus, following a 1992 WHO decision to include HBV in routine vaccination protocols. As a result, although around 6% of people in the region are still living with HBV, most children and teenagers there are protected. But in sub-Saharan Africa, where it's also estimated that about 6% of the population are currently infected, fewer than one-tenth of children receive the necessary inoculations. The region also ranks last in every other intervention, including screening and diagnosis, and in treating those living with the virus.

"Hepatitis B has been, to a large extent, neglected," says Ponsiano Ocamo, a hepatologist at Makerere University in Kampala, Uganda. Health-care workers, he says, are generally under-educated and ill-equipped to treat the virus. Matthews adds that priority for anti-retroviral drugs is weighted so heavily in favour of people with HIV that some health-care workers think those with HBV stand a better chance of receiving adequate care if they contract HIV as well, even though having both infections increases the chance of early death.

With little routine screening, there are also many gaps in researchers' understanding of the prevalence and outcomes of hepatitis in vulnerable populations. While the fight against hepatitis is buoyed by progress in Western Pacific nations, the crisis in sub-Saharan Africa is flying under the radar. "It's a critical time for the region," says Matthews.

KNOWLEDGE GAP

Nuru left her UK screening appointment dejected, and feeling that she knew little about her infection. She turned to the Internet to answer questions she felt had been glossed over by the health-care professionals she saw. Public ignorance about transmission, but awareness that HBV can be passed on during unprotected sex, has given the infection a stigma that, says Nuru, smacks of the whispers that emerged around HIV when that virus first came to light in sub-Saharan Africa. Nuru's body is suppressing the virus well enough that she does not need treatment, but she doesn't talk openly about it. If news that she has HBV spreads back to Uganda, she says, then she worries people will regard her family there with suspicion. "They will be segregated, isolated — they won't get jobs," she says.

Kenneth Kabagambe, who founded Uganda's National Organization for People Living with Hepatitis B (NOPLHB) in 2011, after a friend died with the infection, says he had a similar experience when he himself was diagnosed in 2012. His doctor, he said, left him wondering whether the disease might even be comparable to Ebola.

As Kabagambe and Nuru would learn, hepatitis is sometimes referred to as the silent epidemic, because its carriers do not initially show symptoms. In some cases, the virus responsible can sabotage the liver's function over years without causing noticeable problems, until eventually a viral takeover causes cirrhosis or liver cancer.

Hepatitis C virus (HCV) is an RNA virus that is spread largely through blood — usually through unscreened blood donations, drug use, reuse of unsterilized equipment in hospitals and, to a lesser extent, unprotected sex. There is no vaccine against it, but antiviral medications can cure a chronic infection in most people. By contrast, HBV (a DNA virus, like HIV) is less malignant — in that fewer adults develop chronic infections — but more widespread. It affects almost four times as many people as HCV, and is more likely than HCV to be spread from mother to baby during pregnancy or birth. HBV infection is also divided more along economic lines: it is, says Ocamo, largely "a disease of the poor".

In contrast to people with HIV, adults who don't already have HBV are unlikely to become infected — and, if they do, there is only a small chance of developing a chronic infection or passing it on to other adults.

The group at highest risk of becoming infected and transmitting HBV is infants, who have weaker immune systems. Compared with adults with HBV, toddlers "teem with the virus", says Mark Sonderup, a hepatitis researcher at the University of Cape Town, South Africa. So, screening and treating infected mothers, and vaccinating babies, is key to cracking down on HBV. Yet, myths still circulate among health-workers in Africa about how HBV is transmitted, including that adults with the virus should be isolated. This perpetuates the infection's stigma, says Ocamo.

There are some subtleties to this picture. In Western Pacific nations, the main transmission route for strains of HBV tends to be from mother to baby, according to research¹ that dovetailed with the vaccination campaigns there in the 1990s. In sub-Saharan Africa, however — which has different HBV strains — mothers with the infection tend to have lower viral loads, making it slightly less likely that they will infect their babies during pregnancy or birth. Viral transmission from child to child, through the usual scratches of rough play and the lacklustre hygiene of youth, seems to be a more prominent infection route.

VACCINE PUSH

For many years, policymakers thought that rolling out vaccinations would be enough to halt HBV, says Maud Lemoine, a hepatologist at Imperial College London. That's true in principle, but the vaccine's design makes it difficult to administer. It is generally given in three parts. The first is a 'birth dose', which is most effective if given within 24 hours of birth. The other two doses are given later and several weeks apart. From 1990 to 2015, the proportion of children getting three HBV inoculations skyrocketed from 1% to 84%, with the Western Pacific leading the way at more than 90% coverage, just above that in the Americas; Africa lags behind at 70%².

But in practice, the first dose is not always given at birth — coverage of this dose is only 39% globally — and its timing is not always reported. In Africa, coverage at birth is just 10%. Administering a birth dose within 24 hours, and follow-up vaccinations on schedule, poses a monumental challenge in a region where many births are not supervised by medical professionals.

The challenge of accessing mothers in time has been compounded by a reliance on Gavi, the Vaccine Alliance, an international organization that connects public and private sectors to roll out vaccines. Gavi has been a driving force in expanding HBV vaccination in sub-Saharan Africa. But it does this through a compound inoculation that immunizes against diphtheria, pertussis, tetanus, HBV and influenza, but which isn't given until 6–8 weeks of age. A spokesperson says that the organization has not focused on providing the birth-dose vaccine, in part because it had not seen evidence that distribution systems could get the inoculations to infants within 24 hours of birth, and because it felt the more expensive 5-fold vaccine was a better target for subsidy.

On 29 November, however, Gavi's board voted to prioritize investment in HBV birth-dose vaccines, as part of a strategy targeting six new vaccine programmes from 2021 to 2025. And success in other vaccination campaigns show that it should be possible to overcome distribution challenges. In the 1990s, researchers in Indonesia gave pre-packaged single-use hepatitis B vaccines to local midwives so that they could administer an inoculation after home births, an approach now used more widely³. And two years ago, researchers in Laos demonstrated that providing mobile phones to vigilant health workers and local volunteers helped keep track of births and ensure more infants were vaccinated⁴.

SCREENING RESEARCH

Another key to tackling HBV is screening and diagnosing adults. Mothers are among the most crucial people to check because of their propensity to pass the virus on to their babies. "If you find infected antenatal women, you can also screen their partners. You can vaccinate any household contacts who aren't infected. You can identify any other household contacts who are infected and treat them," says Matthews. "It gives you a route into more population-level interventions."

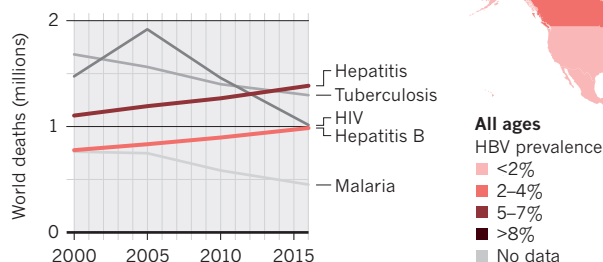
But mothers are not routinely checked before giving birth. Add to that a paucity of cancer registries with accurate data on liver cancer, and a

THE BURDEN OF HEPATITIS B

More than 250 million people live with the virus; few of them are diagnosed and not enough children are vaccinated against it.

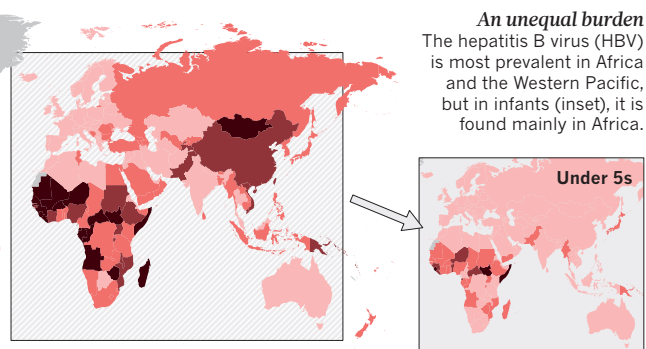
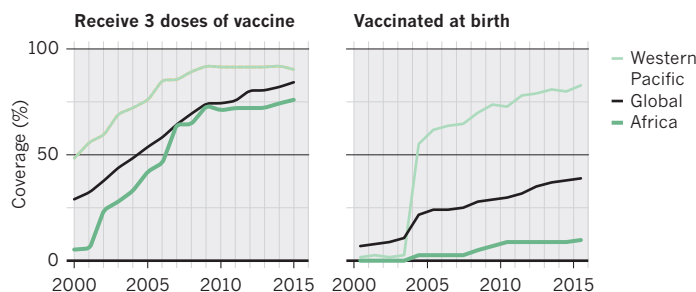
Rising death toll

Hepatitis infections are now associated with more deaths globally than are tuberculosis, HIV or malaria.



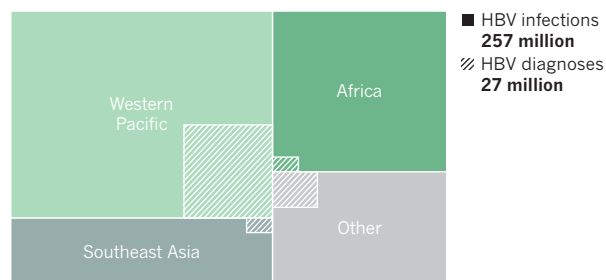
Vaccination lag

Africa is the least-vaccinated region against hepatitis B; the Western Pacific, the most. Only one in ten infants in Africa are vaccinated at birth.



Diagnosis gap

The World Health Organization wants to diagnose 90% of hepatitis B infections by 2030. The rate is currently 10%.



generally low regional turnout for testing, and it's of little surprise that researchers' picture of the prevalence and dynamics of hepatitis viruses are plagued with gaps.

Instead, the populations that are screened most reliably are those who donate blood and people such as Nuru and Kabagambe, who saw firsthand how HIV ravaged their communities, and decided to get tested. Many health professionals have criticized initiatives such as Gavi and the US President's Emergency Plan for AIDS Relief for not doing more to leverage HIV-testing networks to also provide screening for hepatitis. Lemoine points out that one negative HBV test is probably all that an adult needs, because it is so unlikely that they will be infected, whereas people might need to be retested for HIV many times.

Initial screens cost only a few dollars: health workers simply check the person's blood for evidence that their immune system has developed antibodies against the hepatitis viruses. But these checks, says Matthews, test only whether you've been exposed to the viruses, not whether you're currently infected. To get a definitive diagnosis, people need more-expensive nucleic-acid tests that detect the viral DNA of HBV (or, for HCV, viral RNA). The cost can be as high as US\$200 — something that few people in sub-Saharan Africa can afford, says Olufunmilayo Lesi, a member of the WHO's advisory group on viral hepatitis. Fewer than 1% of those in the region with HBV, and 6% of those with HCV, are diagnosed, according to a WHO estimate².

DRIVING FORWARD

Several countries in sub-Saharan Africa are now expanding their screening efforts, including Uganda, which hopes to tie its effort to a vaccination drive aimed at mothers and infants, says Ocama. And researchers have been working on more convenient diagnostic tests. In 2017, the WHO approved a test that detects HCV RNA and runs on equipment found in most hospitals in sub-Saharan Africa — the GeneXpert nucleic-acid system. Made by Cepheid, a company in Sunnyvale, California, it is already used to diagnose HIV and tuberculosis. A test for HBV that could be run on the GeneXpert machine is in beta testing, says Sonderup, but has yet to be formally released. (Cepheid did not reply to requests for comment.)

As the world has focused on combating HIV, billions of dollars have

been poured into developing antiretrovirals — drugs that people with HIV take indefinitely to inhibit the replication of DNA viruses. In low-income countries, the cost of these drugs is heavily subsidized, and in many cases, the same drugs can treat both HIV and HBV.

But when it comes to access to drugs, people with HBV in many resource-limited regions find themselves overlooked in favour of those with HIV. Ocama says he has known hospital administrators who have allowed physicians to administer drugs reserved for people with HIV to those with HBV — but overall, an abysmally small fraction of people in sub-Saharan Africa with HBV receive treatment.

Some countries are increasingly aware that antiretroviral drugs need to also reach people with hepatitis. In 2012, Uganda became the first sub-Saharan African country to produce a generic form of the antiretroviral tenofovir, through the company Quality Chemicals, and the drug is offered for free at some treatment centres, says Ocama. And in 2017, after years of using HIV programmes to secure drugs for people with HBV, the Senegalese Society of Gastroenterology convinced the government to make tenofovir available to them at a price similar to that offered to those with HIV.

Still, the stigma of having HBV can be as problematic as drug scarcity. Patient groups in Africa, Ocama says, are too few and far between. "For many people, I think it is a lonely journey. It is a place of isolation," says Nuru. But she and Kabagambe are determined to change this. After Nuru was diagnosed, she convinced her siblings to get tested. Three out of six tested positive for HBV. Since then, leveraging her sisters in Uganda as part of a 'whisper network', she has convinced 13 other people to be tested, and paid for the procedure.

Meanwhile, the patient network that Kabagambe founded is dedicated to educating the public about HBV and establishing a community in which people who have the virus can talk about it. "Being diagnosed with hepatitis B does not define your end," he says. "You can still prosper." ■

Ian Graber-Stiehl is a science writer in Chicago, Illinois.

1. Gust, I. D. *Gut* **38**, S18–S23 (1996).
2. World Health Organization. *Global Hepatitis Report, 2017* (WHO, 2017).
3. Sutanto, A., Suarnawa, I. M., Nelson, C. M., Stewart, T. & Indijati Soewars, T. *Bull. World Health Organ.* **77**, 119–126 (1999).
4. Xeuvatongsa, A. et al. *Vaccine* **34**, 5777–5784 (2016).

COMMENT

CLIMATE Models have three gaps; brace for faster warming **p.30**

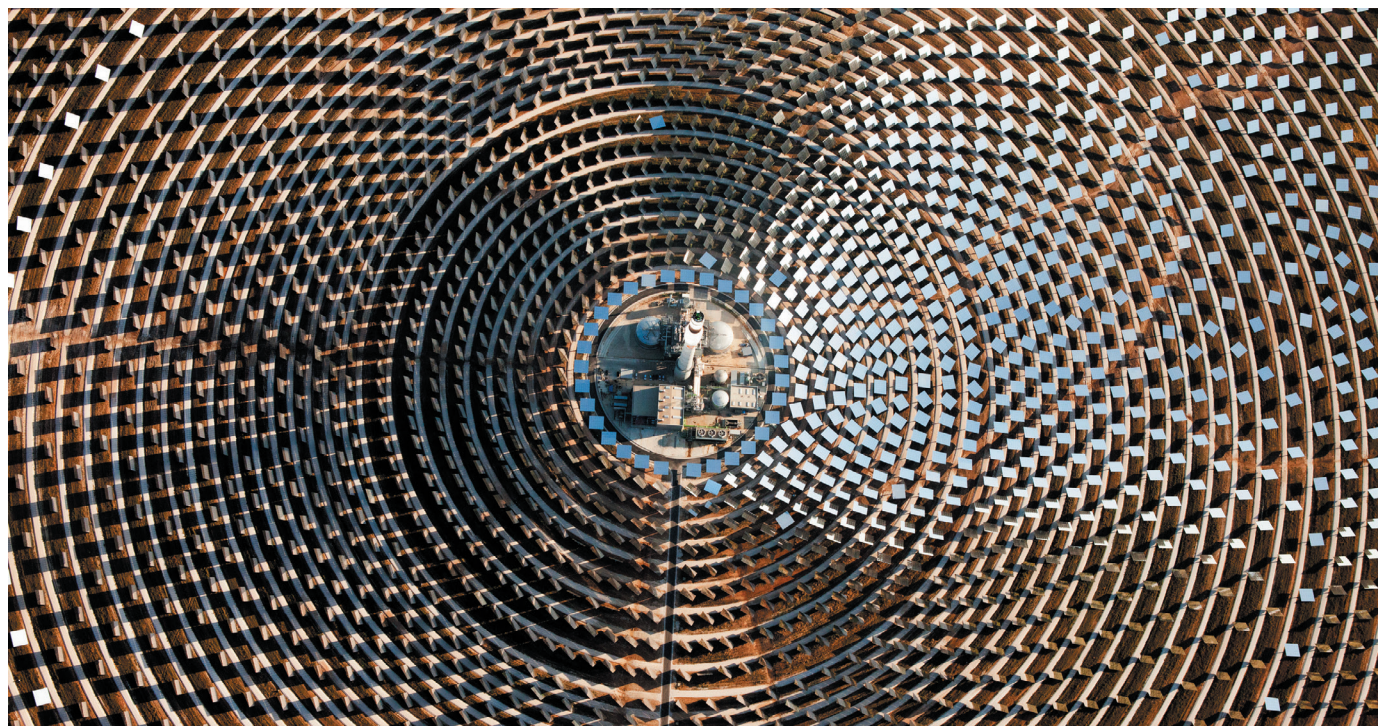


SUSTAINABILITY More carbon in soil is a win-win, for climate and food **p.32**

ARCHITECTURE Where's the proof that swanky labs spark discovery? **p.36**

ROAD PRICING Drivers are unlikely to change the habits of a lifetime **p.39**

MARKEL REDONDO/PANOS



The Gemasolar Thermosolar Plant in Andalusia, Spain.

Emissions are still rising: ramp up the cuts

With sources of renewable energy spreading fast, all sectors can do more to decarbonize the world, argue **Christiana Figueres** and colleagues.

Representatives of 190 nations gather this week to review progress at the annual United Nations climate talks. They face a daunting reality: carbon dioxide emissions from fossil fuels are rising again.

Global CO₂ emissions are projected to go up in 2018 by more than 2% (ref. 1). In 2017, they increased by 1.6%, having flattened out between 2014 and 2016. The reasons? The use of oil and gas keeps growing, and some countries are still using coal to fuel much of their economic growth (see 'Rising pressures').

The UN meetings, this year in Katowice, in the heart of Poland's coalfields, constitute a checkpoint. The Paris climate agreement

was adopted in 2015 — when nations signed up to limit global warming to well below 2 °C, and to strive for 1.5 °C. The first formal revisions of national emissions-reduction targets are in 2020.

To get back on track, the revised targets must be more ambitious than those pledged in 2015. As we argued last year in *Nature*², global CO₂ emissions must start to fall by 2020 if we are to meet the temperature goals of the Paris agreement.

Every year of rising emissions puts economies and the homes, lives and livelihoods of billions of people at risk. It commits us to the effects of climate change for centuries to

come. Already, the terrible impacts of 1 °C of warming above pre-industrial levels are evident. Disasters triggered by weather and climate in 2017 cost the global economy US\$320 billion, and around 10,000 lives were lost (see go.nature.com/2fldcjl). The full costs of 2018's disasters have yet to be tallied — including Typhoon Mangkhut, hurricanes Florence and Michael, and the heatwaves and wildfires that have ravaged swathes of Europe and the United States. These events are likely to contribute to an exponential rise in damages, amounting to some \$2.2 trillion over the past two decades (see go.nature.com/2r2jyy6). ▶

► When it comes to rises in global average temperature, every fraction of a degree matters. A report published in October by the Intergovernmental Panel on Climate Change (IPCC) projected devastating impacts at 2 °C. These include the loss of almost all the world's coral reefs, and extreme, life-threatening heatwaves that could affect more than one-third of the world's population³. Limiting warming to 1.5 °C will significantly lessen those impacts.

So how can we remain optimistic? A low-carbon world is hard to imagine, yet change often follows when we shift our vision of what is possible.

REASONS FOR OPTIMISM

Already, we have achieved things that seemed unimaginable a decade ago. The 2015 Paris agreement is a good example. When the 2009 climate summit in Copenhagen failed to deliver a global framework for addressing climate change, almost everyone thought it was impossible to do so. Yet over the next six years, thousands of people and institutions made the implausible plausible.

The same is true of decarbonizing the economy by 2050. That goal seems far-fetched today because we are anchored in the high-carbon technologies and economic constructs of the twentieth century. But collectively, we are lifting that anchor and charting a course for a different tomorrow. Here's how.

Key technologies are on track. The world is quickly and irrevocably moving towards a clean, cheap and reliable energy system. Over the past decade, the costs of generating solar energy have plummeted by 80%. Morocco, Mexico, Chile and Egypt are producing solar power for 3 US cents or less per kilowatt hour — cheaper than natural gas.

Installations are growing. Today, more than 50% of new capacity for generating electricity is renewable, with wind and solar doubling every 4 years⁴. In developing countries, renewables now account for the majority of all new power generation, a remarkable turnaround from just a decade ago. If these trends continue, renewables will produce half of the world's electricity by 2030.

Coal is being priced out. A record number of US coal-fired power plants will be retired this year, even relatively new ones. In October, the World Bank declined to finance a 500-megawatt coal-fired power plant in Kosovo — the last coal project in the bank's pipeline. The bank's lending rules require it to “go with the lowest cost option, and renewables have now come below the cost of coal”, according to its president (see go.nature.com/2du6mxr).

However, the electricity grid will not be completely transformed until renewables are able to deliver continuous power. Large batteries that can store and smooth out energy supplies are becoming economical faster than expected. For example, a year ago, the

state of South Australia paired a Tesla battery facility with a local wind farm. By storing power for when demand is highest, the system has already repaid nearly one-third of its upfront capital costs of Aus\$90.6 million (US\$65.8 million). The costs of battery storage are expected to halve by 2030 (see go.nature.com/2daiwdt).

By 2040, energy-storage systems should be capable of handling 7% of the world's total installed power capacity (1,000 gigawatts), supporting even more solar and wind installations. Big batteries will spread beyond utilities, into energy storage for rooftop solar panels, for example. This will allow developing regions to leapfrog the need for fossil-fuel power plants and conventional distribution grids, just as mobile phones overtook landlines.

Advances in battery technology are also propelling wide uptake of electric vehicles. A rarity ten years ago, today there are three million globally on our roads. Plug-in car sales were up by 66% in the first half of this year, compared with 2017. Although electric vehicles represent a small fraction of the 80 million cars sold worldwide last year, that will soon change. Norway, France, the United Kingdom, the Netherlands and India have set deadlines for stopping the sale of new cars that are not electric (Norway's is 2025). Most major car manufacturers have announced either a complete shift to electric vehicles or plans for a transition. In 2016, the Organization of the Petroleum Exporting Countries (OPEC) said there would be 46 million electric vehicles by 2040; now it predicts there will be 253 million by that date (see go.nature.com/2qjaaoj).

Abating air pollution is another powerful driver of change. Globally, air pollution contributes to seven million premature deaths every year — from cardiovascular disease, ischaemic heart disease, stroke, chronic obstructive pulmonary disease and lung cancer. People are becoming less tolerant of particulate and noxious-gas emissions from coal plants, factories and cars. China has closed coal-fired power plants in and near

cities and has limited diesel-engine emissions. Pollution levels in Beijing have fallen by 35% over 5 years, but still have a long way to go. India has nine of the world's ten most polluted cities, according to the World Health Organization. The country's target is to reduce air pollution in 100 cities by 20–30% by 2024.

Heavy industry is also evolving. The Energy Transitions Commission announced last month that chemicals, steel and cement can reach net zero emissions by mid-century at a cost of less than 0.5% of global gross domestic product (GDP), with a marginal impact on living standards⁵.

Subnational action is booming. Three years after the Paris agreement, the political landscape has shifted markedly. In some countries, nationalistic impulses are affecting domestic and international policy, threatening the cooperative, multilateral spirit with which the Paris treaty was forged. For example, the federal US government has signalled its intent to withdraw from the agreement — although it cannot formally do so until 2020. Brazil's new administration has expressed doubts about that country's previously ambitious participation. Climate-sceptic voices have re-emerged in Australia.

Yet support for climate action remains strong in cities, regional governments and the private sector. Globally, more than 9,000 cities and municipalities from 128 countries, representing 16% of the world's population, have reiterated their commitment to the Paris agreement through the Global Covenant of Mayors. So have 245 state and regional bodies from 42 countries, which are home to 17.5% of the global population⁶. Most US citizens live in a jurisdiction that still supports the Paris goals. If all of these US cities, states and companies stick to their emissions-reduction pledges, they could put the country within striking distance of the Paris commitment made by the Obama administration, irrespective of current federal action.



A rider swaps scooter batteries at a roadside facility in Taiwan.

CHRIS STOWERS/PANOS

Boards of directors, presidents of central banks, investors and insurers are increasingly concerned about the economic risks of climate change and the threat to health, water, land and biodiversity resources worldwide. As many as 6,225 companies headquartered in 120 countries have pledged to contribute to the Paris goals⁶, representing \$36.5 trillion in revenue — more than the combined GDP of the United States and China (see go.nature.com/2aphgjs). These firms understand that the agreement is likely to bring \$26 trillion in economic benefits by 2030, including 65 million jobs in the booming low-carbon economy.

Carbon pricing is on the rise. In 2017, 1,400 multinational companies were factoring a price on carbon pollution into their business plans — an eightfold rise since 2014 (see go.nature.com/2p9osnb). Nearly one-quarter of the 155 companies that have committed to switching to solely renewable energy through the RE100 initiative are already getting 95% of their power from clean sources. And close to 500 companies have recognized the business opportunity of setting 'science-based targets' for their emissions reductions.

In 2017, more than 100 members of the Compact of States and Regions reported average reductions in emissions of 8.5%. Twelve members achieved a 20% reduction or more (see go.nature.com/2aphgjs). Global emissions could be cut by one-third by 2030 compared to current national policy pathways, if ambitious initiatives such as the Under2 Coalition, RE100, C40 and the Global Covenant of Mayors all meet their goals.

Bolder Paris targets. Whereas some parties to the Paris agreement are backsliding, many others are signalling their intention to be more ambitious. The agreement's five-year cycle enables a gradual ratcheting up of effort, and bolder targets will be easier to achieve thanks to the market forces noted previously.

China, India and the European Union are setting the pace. These regions represent 40% of global carbon emissions. They are set to achieve more than they agreed in the first round. Their leaders can step up and announce even bolder programmes at the UN summit to review the Paris commitments in September 2019.

Leaders of smaller countries with big plans for a safer climate can do the same. In November, the Marshall Islands, a member of the Climate Vulnerable Forum, became the first country to submit a new, more ambitious climate target to the UN.

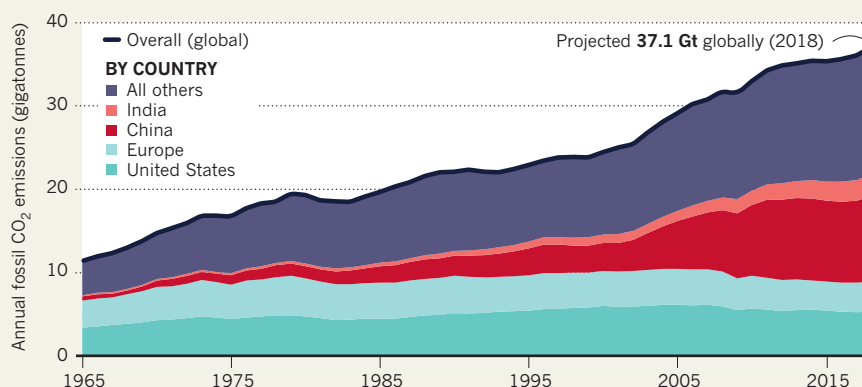
Twenty-two other countries, from Argentina to the United Kingdom, have declared that they will explore the possibility of strengthening their Paris pledges before 2020. Chile announced in February that it would phase out coal completely. And this September, four nations joined the Carbon Neutrality Coalition, bringing the total to 19. Including

RISE IN PRESSURES

Carbon dioxide emissions are growing again after pausing for a few years. Renewable sources of power are just beginning to replace fossil fuels, as their costs become competitive.

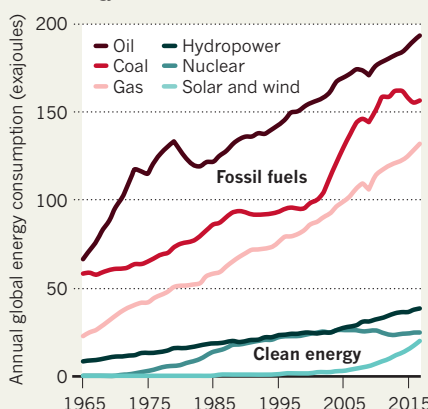
EMISSIONS ARE STILL INCREASING

China and India still rely heavily on coal; the United States and the European Union are slowly decarbonizing.



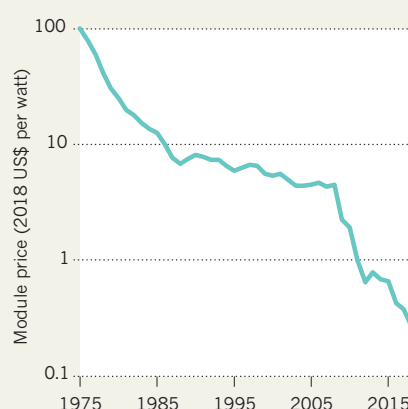
RENEWABLES ARE PICKING UP

Half of all new energy-generation capacity is renewable. Switching to electric cars would prioritize clean energy over oil.



SOLAR ENERGY IS AFFORDABLE

Costs have dropped by 80% over the past decade as solar installations have expanded.



Canada, Mexico, Ethiopia and many in Europe, the coalition's members commit to developing long-term emissions-reduction strategies by 2020.

THE TASK AHEAD

Rising emissions are of grave concern. But the low-carbon transition is snowballing, bowled along by the underlying economics. The task now for ministers and governments meeting in Katowice is to accelerate that momentum and keep everyone on board. This includes completing the rulebook for implementing the Paris agreement so that there is a clear path forward.

All efforts must strive to keep global warming to 1.5°C. Now that we understand the different impacts of a 1.5°C and a 2°C rise, we cannot in good conscience provoke unnecessary risks to the most vulnerable and to the global economy.

First and foremost, the world must quickly replace coal and other fossil fuels with renewables. It is an economic imperative and an ecological necessity. No new investments

should be allocated to expanding fossil-fuel assets, and plans must be made for retiring existing infrastructure as clean technologies take hold.

Leaders must broaden the scope of their actions. The EU intends to ramp up its emissions-reduction goals to 55% by 2030. This ambitious target can be attained by improving energy efficiency, by deploying renewables faster and by a quicker exit from coal. The EU can encourage stricter vehicle-emissions standards and hasten the uptake of electric vehicles, and develop public and shared transport. As one of the world's largest importers of commodities linked to deforestation (such as palm oil), the EU can develop an initiative to combat the loss of forests and other carbon sinks. This would build on a November statement from the French government, which aims to stop imports of non-sustainable forest or agricultural products by 2030.

China can produce a low-emissions development strategy up to 2050, including a plan to replace coal that will ensure its emissions peak before 2030 and at a lower level

than previously planned. China can also ensure that its investments made beyond its borders through the Belt and Road Initiative support renewable energy and protect and restore tropical forests and other sensitive ecosystems.

China and India have made substantial progress with reforestation and have the potential to do more: further tree planting could remove 1.25 gigatonnes and 520 megatonnes of CO₂ per year in each country, respectively.

India can continue to deploy solar farms, leveraging its leadership of the International Solar Alliance to displace coal and clean up its smog-choked cities. By 2020, India can announce its own fossil-fuel exit strategy and a target date for its peak CO₂ emissions.

A shared purpose across all political, civil and industrial sectors is key, as the breadth of authors and co-signatories to this article attests (see go.nature.com/2riswcr for co-signatories). What seemed radical in 2015 is now advantageous. Let us ensure that the exponential curve of solutions outpaces that of climate impacts, and drives net emissions to zero by 2050. It's necessary, desirable and achievable. ■

Christiana Figueres is convenor of Mission 2020 and vice-chair of the Global Covenant of Mayors. **Corinne Le Quéré** is director of the Tyndall Centre for Climate Change Research, University of East Anglia, Norwich, UK. **Anand Mahindra** is chair of the Mahindra Group, Mumbai, India. **Oliver Bäte** is chair of the board of management of Allianz SE, Munich, Germany. **Gail Whiteman** is professor and director of the Pentland Centre for Sustainability in Business, Lancaster University, Bailrigg, Lancaster, UK. **Glen Peters** is research director at the Center for International Climate Research, Oslo, Norway. **Dabo Guan** is chair professor in climate-change economics at the University of East Anglia, Norwich, UK, and distinguished professor at Tsinghua University, Beijing, China. e-mail: cfigueres@mission2020.global

1. Le Quéré, C. et al. *Earth Syst. Sci. Data* <https://doi.org/10.5194/essd-10-2141-2018> (2018).
2. Figueres, C. et al. *Nature* **546**, 593–595 (2017).
3. Intergovernmental Panel on Climate Change. *Global Warming of 1.5 °C* (IPCC, 2018).
4. Global Climate Action Summit. *Exponential Climate Action Roadmap* (Future Earth/Sitira, 2018).
5. Energy Transitions Commission. *Mission Possible: Reaching Net-Zero Carbon Emissions from Harder-To-Abate Sectors by Mid-Century* (ETC, 2018).
6. Yale Data-Driven Environmental Solutions Group. *Who's Acting On Climate Change? Subnational and Non-State Global Climate Action* (Yale Data-Driven, 2017).

A list of co-signatories accompanies this article online (see go.nature.com/2riswcr).



Devastating wildfires ravaged California last month.

Global warming will happen faster than we think

Three trends will combine to hasten it, warn **Yangyang Xu, Veerabhadran Ramanathan and David G. Victor.**

Prepare for the “new abnormal”. That was what California Governor Jerry Brown told reporters last month, commenting on the deadly wildfires that have plagued the state this year. He’s right. California’s latest crisis builds on years of record-breaking droughts and heatwaves. The rest of the world, too, has had more than its fair share of extreme weather in 2018. The *Lancet* Countdown on health and climate change announced last week that 157 million more people were exposed to heatwave events in 2017, compared with 2000.

Such environmental disasters will only intensify. Governments, rightly, want to know what to do. Yet the climate-science community is struggling to offer useful answers.

In October, the Intergovernmental Panel on Climate Change (IPCC) released a report setting out why we must stop global

warming at 1.5 °C above pre-industrial levels, and how to do so¹. It is grim reading. If the planet warms by 2 °C — the widely touted temperature limit in the 2015 Paris climate agreement — twice as many people will face water scarcity than if warming is limited to 1.5 °C. That extra warming will also expose more than 1.5 billion people to deadly heat extremes, and hundreds of millions of individuals to vector-borne diseases such as malaria, among other harms.

But the latest IPCC special report underplays another alarming fact: global warming is accelerating. Three trends — rising emissions, declining air pollution and natural climate cycles — will combine over the next 20 years to make climate change faster and more furious than anticipated. In our view, there’s a good chance that we could breach the 1.5 °C level by 2030, not by 2040 as projected



GENE BLEVINS/REUTERS

in the special report (see ‘Accelerated warming’). The climate-modelling community has not grappled enough with the rapid changes that policymakers care most about, preferring to focus on longer-term trends and equilibria.

Policymakers have less time to respond than they thought. Governments need to invest even more urgently in schemes that protect homes from floods and fires and help people to manage heat stress (especially older individuals and those living in poverty). Nations need to make their forests and farms more resilient to droughts, and prepare coasts for inundation. Rapid warming will create a greater need for emissions policies that yield the quickest changes in climate, such as controls on soot, methane and hydrofluorocarbon (HFC) gases. There might even be a case for solar geoengineering — cooling the planet by, for instance, seeding reflective particles in the stratosphere to act as a sunshade.

Climate scientists must supply the evidence policymakers will need and provide assessments for the next 25 years. They should advise policymakers on which climate-warming pollutants to limit first to gain the most climate benefit. They should assess which policies can be enacted most swiftly and successfully in the real world, where political, administrative and economic constraints often make abstract, ‘ideal’ policies impractical.

SPEEDING FREIGHT TRAIN

Three lines of evidence suggest that global warming will be faster than projected in the recent IPCC special report.

First, greenhouse-gas emissions are still

rising. In 2017, industrial carbon dioxide emissions are estimated to have reached about 37 gigatonnes². This puts them on track with the highest emissions trajectory the IPCC has modelled so far. This dark news means that the next 25 years are poised to warm at a rate of 0.25–0.32°C per decade³. That is faster than the 0.2°C per decade that we have experienced since the 2000s, and which the IPCC used in its special report.

Second, governments are cleaning up air pollution faster than the IPCC and most climate modellers have assumed. For example, China reduced sulfur dioxide emissions from its power plants by 7–14% between 2014 and 2016 (ref. 4). Mainstream climate models had expected them to rise. Lower pollution is better for crops and public health⁵. But aerosols, including sulfates, nitrates and organic compounds, reflect sunlight. This shield of aerosols has kept the planet cooler, possibly by as much as 0.7°C globally⁶.

Third, there are signs that the planet might be entering a natural warm phase that could last for a couple of decades. The Pacific Ocean seems to be warming up, in accord with a slow climate cycle known as the Interdecadal Pacific Oscillation⁷. This cycle modulates temperatures over the equatorial Pacific and over North America. Similarly, the mixing of deep and surface waters in the Atlantic Ocean (the Atlantic meridional overturning circulation) looks to have weakened since 2004, on the basis of data from drifting floats that probe the deep ocean⁸. Without this mixing, more heat will stay in the atmosphere rather than going into the deep oceans, as it has in the past.

These three forces reinforce each other. We estimate that rising greenhouse-gas emissions, along with declines in air pollution, bring forward the estimated date of 1.5°C of warming to around 2030, with the

2°C boundary reached by 2045. These could happen sooner with quicker shedding of air pollutants. Adding in natural decadal fluctuations raises the odds of blasting through 1.5°C by 2025 to at least 10% (ref. 9). By comparison, the IPCC assigned probabilities of 17% and 83% for crossing the 1.5°C mark by 2030 and 2052, respectively.

FOUR FRONTS

Scientists and policymakers must rethink their roles, objectives and approaches on four fronts.

Assess science in the near term. Policymakers should ask the IPCC for another special report, this time on the rates of climate change over the next 25 years. The panel should also look beyond the physical science itself and assess the speed at which political systems can respond, taking into account pressures to maintain the status quo from interest groups and bureaucrats. Researchers should improve climate models to describe the next 25 years in more detail, including the latest data on the state of the oceans and atmosphere, as well as natural cycles. They should do more to quantify the odds and impacts of extreme events. The evidence will be hard to muster, but it will be more useful in assessing real climate dangers and responses.

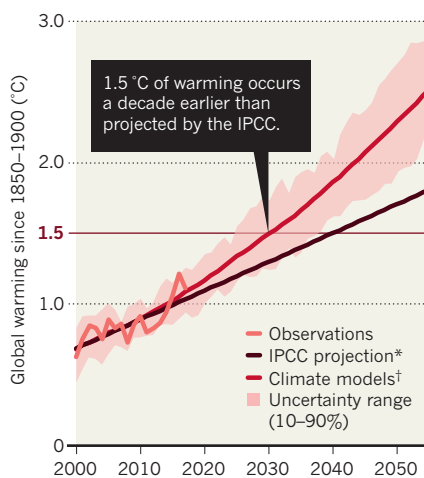
Rethink policy goals. Warming limits, such as the 1.5°C goal, should be recognized as broad planning tools. Too often they are misconstrued as physical thresholds around which to design policies. The excessive reliance on ‘negative emissions technologies’ (that take up CO₂) in the IPCC special report shows that it becomes harder to envision realistic policies the closer the world gets to such limits. It’s easy to bend models on paper, but much harder to implement real policies that work.

Realistic goals should be set based on political and social trade-offs, not just on geophysical parameters. They should come out of analyses of costs, benefits and feasibility. Assessments of these trade-offs must be embedded in the Paris climate process, which needs a stronger compass to guide its evaluations of how realistic policies affect emissions. Better assessment can motivate action but will also be politically controversial: it will highlight gaps between what countries say they will do to control emissions, and what needs to be achieved collectively to limit warming. Information about trade-offs must therefore come from outside the formal intergovernmental process — from national academies of sciences, subnational partnerships and non-governmental organizations.

Design strategies for adaptation. The time for rapid adaptation has arrived. Policymakers need two types of information from scientists to guide their responses.

ACCELERATED WARMING

Climate simulations predict that global warming will rise exponentially if emissions go unchecked.



*Trend for 2001–15 extended with a constant rate of 0.2°C per decade, as per IPCC special report. †Ten-year average, 37 climate models for the RCP8.5 scenario (IPCC Fifth Assessment, 2014).

SOURCES: REF. 1/GISTEMP/IPCC FIFTH ASSESSMENT REPORT (2014)

First, they need to know what the potential local impacts will be at the scales of counties to cities. Some of this information could be gleaned by combining fine-resolution climate impact assessments with artificial intelligence for 'big data' analyses of weather extremes, health, property damage and other variables. Second, policymakers need to understand uncertainties in the ranges of probable climate impacts and responses. Even regions that are proactive in setting adaptation policies, such as California, lack information about the ever-changing risks of extreme warming, fires and rising seas. Research must be integrated across fields and stakeholders — urban planners, public-health management, agriculture and ecosystem services. Adaptation strategies should be adjustable if impacts unfold differently. More planning and costing is needed around the worst-case outcomes.

Understand options for rapid response. Climate assessments must evaluate quick ways of lessening climate impacts, such as through reducing emissions of methane, soot (or black carbon) and HFCs. Per tonne, these three 'super pollutants' have 25 to thousands of times the impact of CO₂. Their atmospheric lifetimes are short — in the range of weeks (for soot) to about a decade (for methane and HFCs). Slashing these pollutants would potentially halve the warming trend over the next 25 years¹⁰.

There has been progress on this front. At

the Global Climate Action summit held in September in San Francisco, California, the United States Climate Alliance — a coalition of state governors representing 40% of the US population — issued a road map to reduce emissions of methane, HFCs and soot by 40–50% by 2030 (see go.nature.com/2ozhojc). The 2016 Kigali amendment to the Montreal

“More planning and costing is needed around the worst-case outcomes.”

Protocol, which will go into force by January 2019, is set to slash HFC emissions by 80% over the next 30 years. Various climate engineering options should be on the table as an emergency response. If global conditions really deteriorate, we might be forced to extract large volumes of excess CO₂ directly from the atmosphere. An even faster emergency response could be to inject aerosols into the atmosphere to lower the amount of solar radiation heating the planet, as air pollution does. This option is hugely controversial, and might have unintended consequences, such as altering rainfall patterns that lead to drying of the tropics. So research and planning are crucial, in case this option is needed. Until there is investment in testing and technical preparedness — today, there is almost none — the chances are high that the wrong kinds of climate-engineering scheme will be deployed by irresponsible parties who are uninformed by research¹¹.

For decades, scientists and policymakers

have framed the climate-policy debate in a simple way: scientists analyse long-term goals, and policymakers pretend to honour them. Those days are over. Serious climate policy must focus more on the near-term and on feasibility. It must consider the full range of options, even though some are uncomfortable and freighted with risk. ■

Yangyang Xu is an assistant professor of atmospheric sciences at Texas A&M University, College Station, Texas, USA. **Veerabhadran Ramanathan** is professor of atmospheric and climate sciences and **David G. Victor** is professor of international relations at the University of California, San Diego, USA.
e-mail: david.victor@ucsd.edu

1. Intergovernmental Panel on Climate Change. *Global Warming of 1.5 °C* (IPCC, 2018).
2. Le Quéré, C. et al. *Earth Syst. Sci. Data* **10**, 405–448 (2018).
3. Smith, D. M. et al. *Geophys. Res. Lett.* <https://doi.org/10.1029/2018GL079362> (2018).
4. Karplus, V. J., Zhang, S. & Almond, D. *Proc. Natl Acad. Sci. USA* **115**, 7004–7009 (2018).
5. Burnett, R. et al. *Proc. Natl Acad. Sci. USA* **115**, 9592–9597 (2018).
6. Salzmann, M. *Sci. Adv.* **2**, e1501572 (2016).
7. Meehl, G. A., Hu, A. & Teng, H. *Nature Commun.* **7**, 11718 (2016).
8. Chen, X. & Tung, K.-K. *Nature* **559**, 387–391 (2018).
9. Henley, B. J. & King, A. D. *Geophys. Res. Lett.* **44**, 4256–4262 (2017).
10. Xu, Y. & Ramanathan, V. *Proc. Natl Acad. Sci. USA* **114**, 10315–10323 (2017).
11. Victor, D. G. *Oxford Rev. Econ. Pol.* **24**, 322–336 (2008).

Put more carbon in soils to meet Paris climate pledges

Take these eight steps to make soils more resilient to drought, produce more food and store emissions, urge **Cornelia Rumpel** and colleagues.

Soils are crucial to managing climate change. They contain two to three times more carbon than the atmosphere. Plants circulate carbon dioxide from the air to soils, and consume about one-third of the CO₂ that humans produce. Of that, about 10–15% ends up in the earth.

Carbon is also essential for soil fertility and agriculture. Decomposing plants, bacteria, fungi and soil fauna, such as earthworms, release organic matter and nutrients for plant growth, including nitrogen and phosphorus. This gives structure to soil, making it resilient to erosion and able to hold water. Typically, organic matter accounts for a few per cent

of the mass of soil near the surface.

Increasing the carbon content of the world's soils by just a few parts per thousand (0.4%) each year would remove an amount of CO₂ from the atmosphere equivalent to the fossil-fuel emissions of the European Union¹ (around 3–4 gigatonnes (Gt)). It would also boost soil health: in studies across Africa, Asia and Latin America, increasing soil carbon by 0.4% each year enhanced crop yields by 1.3% (ref. 2).

Yet one-third of the world's soils are degraded³. Poor farming practices, industry and urbanization take their toll. Throughout human history, 133 Gt of carbon have been lost from soils, adding almost 500 Gt

of CO₂ to the atmosphere⁴. As the amount of organic matter dwindles, soils face mounting damage from erosion, heatwaves and droughts — it is a vicious circle. In the worst cases, nothing can be grown. This is what happened in the 1930s 'dust bowl' in the central southern United States.

Improving soil carbon is now high on the political agenda. In 2015 at the Paris climate summit, France launched the 4p1000 initiative — to promote research and actions globally to increase soil carbon stocks by 4 parts per 1,000 per year. We are members of the scientific and technical committee for this initiative.

In November 2017 at the Bonn



Air pollution in Sumatra in 2013, where peatlands were burned to clear land for a palm-oil plantation.

ULET INFANSASTI/GREENPEACE

climate conference in Germany, delegates established the Koronivia Joint Work on Agriculture programme. Tasked with helping farmers to reduce emissions and maintain food security in a changing climate, it will hold its first workshop this week at the annual summit of the United Nations Framework Convention on Climate Change (UNFCCC) in Katowice, Poland.

We call on countries involved in the Koronivia process to establish a body to monitor soil carbon in farmland, map changes to it and reclaim degraded areas. All involved should focus on the eight steps set out below.

EIGHT STEPS

The following practices would increase the amount of carbon held globally in soil:

Stop carbon loss. Protecting peatlands is the first priority for keeping existing carbon in the ground. These hold between 32% and 46% of all soil carbon (an estimated 500–700 Gt of approximately 1500 Gt) in an area about half the size of Brazil. Each year they take up about 1% of the global CO₂ emissions generated by humans⁵.

Yet 10–20% of peatlands have been drained or burned and converted to agriculture, particularly in tropical areas. For example, fires used to clear land in maritime southeast Asia blanketed much of Indonesia in a toxic yellow haze during September and October 2015, emitting more CO₂

per day than the whole of the European Union. Globally, such destruction is using up 1–2 Gt CO₂ per year of the remaining emissions budget necessary to stay within the Paris climate targets. To protect this resource, governments must ban burning of peatlands, stop their use in agriculture, or plan and enforce practices that preserve peat through continuous wet conditions.

Degraded mineral soils also need to be restored by controlling grazing, applying green manure or growing cover crops. Between 10 million and 60 million square kilometres of soils are degraded — up to 40% of the world's land area⁶. Restored, these could take up 9–19% of global CO₂ emissions for 25–50 years, at rates of 3–7 Gt of CO₂ per year.

One global effort is making a start. The Bonn Challenge aims to improve 1.5 million km² of degraded and deforested land by 2020 (and 3.5 million km² by 2030) through conservation, recovery and sustainable management of forests and other ecosystems. It is overseen by the Global Partnership on Forest and Landscape Restoration, and is run by the International Union for Conservation of Nature.

Promote carbon uptake. Researchers need to establish a set of best practices for getting more carbon into soil. Proven techniques include making sure the soil is planted all year round, adding crop residues such as mulch and straw or compost,

and minimizing tillage practices such as ploughing. In areas at high risk of erosion, contour farming and terracing should be implemented. Agroforestry systems, hedges and wetlands can increase biodiversity and soil carbon. Planting nitrogen-fixing plants such as beans, alfalfa and oilseed rape reduces the need for mineral fertilizers, which can release nitrous oxide, a greenhouse gas that is around 300 times more potent than CO₂ (ref. 7).

Soils need regular inputs of organic matter. Competing demands for crop residues (also used as fodder) or dung (also used in cooking or heating) can limit what is available. Shortages of other soil nutrients might reduce the capacity of plants to produce enough organic matter to restore all soils.

Regional strategies for increasing soil carbon need to be developed, taking into account local soil types, climates, rates of climate change and socioeconomic contexts. These will favour particular plant species and restrict certain practices. For example, burning stubble or straw for land clearance should be prevented in Asia and South America. Similarly, slash and burn of tropical forests should be avoided in Africa. In Europe, reducing mineral fertilizers and implementing agroecological practices would be effective.

Monitor, report and verify impacts.

Researchers and land managers need to track and evaluate interventions. Large-scale,

long-term frequent monitoring is costly. It involves extensive field surveys that collect hundreds of samples per hectare, with laboratory analyses costing up to US\$10 per sample. And to yield sufficient georeferenced data to capture small changes in soil organic carbon over time, it must continue for at least 10 years. Obtaining access to private land is one challenge. Another is a lack of technical expertise and knowledge, especially in developing countries.

The Global Soil Laboratory Network (GLOSOLAN) is working to improve matters by harmonizing protocols and standards and setting up global training programmes in soil analysis. GLOSOLAN is part of the Global Soil Partnership run by the Food and Agriculture Organization of the United Nations.

Deploy technology. Advanced instruments make soil measurements cheaper, faster and more accurate. Portable infrared spectrometers will soon be capable of tracking multiple chemical signatures in soil, including carbon, for less than \$1 per sample. Harmonized methodologies, verification standards and common guidelines will be needed for all these devices. Satellite imagery is also essential for scanning wide areas. Researchers should design automatic procedures and algorithms for assessing soil carbon content from space, or for predicting it from the characteristics of vegetation. These techniques should work whether soils are wet or dry, and for surfaces that are rough or smooth. They will require rigorous, ground-based verification.

Test strategies. Computer models and a network of field sites⁸ need to be developed to test the effectiveness of, say, avoiding ploughing. Farms should report their actions, verified by spot checks, field surveys or remote sensing. Data on soil types and meteorological variables will also need to be collected. Some open research databases exist: the Integrated Carbon Observation System measures exchanges of greenhouse gases across 120 experimental sites in Europe, for instance. But soil data need to be more transparent and accessible.

Involve communities. The public should be made more aware of the importance of soil organic carbon and of their ability to improve it on farms, in private gardens and public areas. Citizen-science approaches to collecting data, which are widely used in urban planning, for example, should be extended to soils. A good example is the earthworm population survey conducted by farmers across 1,300 hectares in the United Kingdom, which helped to assess farmland biodiversity (see also J. L. Stroud *Nature* **562**, 344; 2018).

A global, open, online platform to collect and share soil carbon data needs to be established. It could be based on GlobalSoil-Map, which was set up by scientists in 2009.

Basing it on a widely used technology, such as a geographic information system (GIS), would broaden its reach and reduce the need for training. Such open platforms will be important in developing countries, where access to resources is limited.

Coordinate policies. Political frameworks covering soils and climate change should work together. These include parties involved in SDG15 — the UN Sustainable Development Goal that seeks to halt and reverse land degradation by 2030 — and the UN Convention to Combat Desertification, which has targets and funding for stopping land degradation and managing land sustainably. Scientists should help countries to integrate soil carbon goals in their pledged emissions cuts to the Paris agreement. And the Koronivia programme should develop complementary targets for storing soil carbon.

Targets and policies will be needed to reform agricultural practices worldwide, which will take decades. Farmers will need incentives to change their methods. Financial compensation could be given to cover costs and risks, for example. Researchers need a better understanding of geographical priorities, such as hotspots that combine harsh climates and vulnerable populations.

Provide support. Policymakers should include soil carbon in emissions-trading schemes and carbon taxes. This will be harder than schemes for CO₂ because soil carbon is transient, unevenly distributed and harder to measure. Crop insurance and other services can offer premiums to farmers who have improved soil carbon. Carbon credits or discounts could be given for lands that are at risk of soil-carbon loss⁹.

Some governments have begun to act. India has distributed soil-health cards to 100 million farmers. These explain how to test soil for nutrients and choose fertilizers. China has banned agricultural fires and subsidizes farmers who return residues to fields¹⁰. The United States compensates farmers who remove cropland from production and increase areas of carbon-rich grasslands.

Development banks and investors should create global investment funds to support practices that improve soil carbon. These could be similar to the Moringa fund, which targets agroforestry projects in Latin America and sub-Saharan Africa.

WHAT NEXT?

First, researchers, policymakers and land managers need to recognize that increasing soil carbon stocks and protecting

carbon-rich soils is crucial for achieving the Paris climate targets and SDGs. Policy-focused organizations should convene a joint forum to coordinate action. This could be hosted by the 4p1000 initiative. Neighbouring countries should exchange experiences, develop common management strategies and make joint decisions on climate-change mitigation, adaptation and land degradation.

Second, international funding agencies should set up a pool of several million dollars to address urgent research gaps, such as those identified by the 4p1000 initiative. These include: estimating the potential for soil carbon storage; developing targets and management practices; designing monitoring, reporting and verification strategies; and understanding basic soil-plant processes.

As the Koronivia summit begins, governments must pledge funds to bring together soil experts, donors and policymakers to act on soil carbon storage. ■

Cornelia Rumpel is chair of the 4p1000 initiative's scientific and technical committee and director of research at the CNRS Institute of Ecology and Environmental Sciences, Thiverval-Grignon, France. **Farshad Amiraslani** is deputy dean of academic affairs, Faculty of Geography, University of Tehran, Iran. **Lydie-Stella Koutika** is a senior researcher and deputy director of the Research Center on Productivity and Sustainability of Industrial Plantations (CRDPI), Pointe-Noire, Republic of the Congo. **Pete Smith** is professor of soils and global change, University of Aberdeen, UK. **David Whitehead** is a plant and soil scientist at Manaaki Whenua — Landcare Research, Lincoln, New Zealand. **Eva Wollenberg** is a research professor and flagship leader in the CGIAR Research Program CCAFS, University of Vermont, Burlington, USA. e-mail: cornelia.rumpel@inra.fr

1. Chabbi, A. et al. *Nature Clim. Change* **7**, 307–309 (2017).
2. Soussana, J.-F. et al. *Soil Tillage Res.* <https://doi.org/10.1016/j.still.2017.12.002> (2017).
3. Food and Agriculture Organization of the United Nations and Intergovernmental Technical Panel on Soils. *Status of the World's Soil Resources — Main Report* (FAO & ITPS, 2015).
4. Sanderman, J., Hengl, T. & Fiske, G. J. *Proc. Natl Acad. Sci. USA* **114**, 9575–9580 (2017).
5. Leifeld, J. & Menichetti, L. *Nature Commun.* **9**, 1071 (2018).
6. Gibbs, H. K. & Salmon, J. M. *Appl. Geogr.* **57**, 12–21 (2015).
7. IPCC. *Guidelines for National Greenhouse Gas Inventories. Volume 4: Agriculture, Forestry and Other Land Use* (IPCC, 2006).
8. Smith, P. et al. *Glob. Change Biol.* **18**, 2089–2101 (2012).
9. Thamo, T. & Pannell, D. J. *Clim. Pol.* **16**, 973–992 (2016).
10. Zhao, Y. et al. *Proc. Natl Acad. Sci. USA* **115**, 4045–4050 (2018).

A list of co-signatories accompanies this article online (see go.nature.com/2skatbf).



The bold design of the Salk Institute for Biomedical Studies in La Jolla, California, is intended to attract star scientists.

COMMUNITY

How luxe is your laboratory?

Kendall Powell probes a study claiming that swanky architecture sparks discovery.

From the Francis Crick Institute in London to Japan's Okinawa Institute of Science and Technology, much has been made of how architecture influences scientists' work. That is, how sunlit benches help researchers' mental health; how cushy breakout spaces spark spontaneous collaboration; and how walking trails and yoga classes rebalance workaholic tendencies. Indeed, many who work in academic and corporate science agree that built-in amenities add productivity.

As someone who has chronicled scientists' lives for *Nature* and other media outlets for nearly two decades, I've heard a great deal about the power of place to boost or sap the will of postdocs and principal investigators. But, as a former

Laboratory Lifestyles: The Construction of Scientific Fictions

SANDRA KAJI-O'GRADY,
CHRIS L. SMITH AND
RUSSELL HUGHES
MIT Press (2019)

sometimes surprising journey through the history and trends of laboratories built around lifestyle — scientists' conversations, proclivities and interactions, not just their apparatus. Edited by Australia-based architecture scholars Sandra Kaji-O'Grady, Chris Smith and Russell Hughes, the journey begins in the 1950s and 1960s in California, then, as now, a magnet

cell biologist, I want to see the data. So I picked up *Laboratory Lifestyles* with some anticipation.

What I found was a book that, although not strong on data, offers an agreeable,

for science. An early chapter showcases how the southern Californian lifestyle of surfing and outdoor living crept into the design of the RAND Corporation's original 'waffle' building in Santa Monica, and the sweeping ocean-to-mountain vistas of the Hughes Research Laboratories in Malibu. The 'work hard, play hard' mantra guided coastal California's deep thinkers long before biotechnology company Genentech and its amenity-fuelled approach to research arrived in South San Francisco.

OPEN-PLAN INNOVATION

The book rightly dwells on the architectural breakthrough of Louis Kahn's 1963 Salk Institute for Biological Studies in

ANDRIY BLOKHIN/ALAMY

La Jolla, California, with its imposing concrete facades, teak accents and white travertine marble courtyard. That bold facade was intended to lure star scientists, philanthropists and partners, and engage the public. It has done all this. When I was a graduate student there, the views of paragliders over the Pacific Ocean and the sea breezes lifted my spirits amid the worst experimental fails. What I did not

“The lack of evidence that the hipster-hub aesthetic actually recruits, retains or spurs innovators is alarming.”

understand then was the Salk’s real breakthrough: its open-plan lab benches, crafted to encourage conversations and enable easy rearrangements as science evolved. Soon, this innovation was adopted the world over. (The book does not cover more-controversial aspects of the Salk’s configuration: its separation of senior and junior staff, for instance, has been criticized as elitist.)

Contributors Kathleen Brandt and Brian Lonsway take us to the early 1970s with Xerox’s Palo Alto Research Center (PARC) conference room, a haven decked with then-novel beanbag chairs and whiteboard walls instead of a conference table. Set in the then-nascent Silicon Valley, PARC’s output was attributed as much to its culture as to the talent it attracted. Its ‘creative hive’ atmosphere has since been recreated, with heavy investment, at workplaces ranging from Google to biotech up-and-comer Moderna Therapeutics in Cambridge, Massachusetts. But did the beanbags boost productivity? The authors write that it is “impossible to prove causality”. Given that establishing causalities is scientists’ lifeblood, the lack of evidence that the hipster-hub aesthetic actually recruits, retains or spurs innovators is alarming.

In the 2000s, eminent architects created lab buildings for two companies in Basel, Switzerland — Actelion (designed by Herzog and de Meuron) and Novartis (Frank Gehry, among others) — along with Singapore’s science-hub campus one-north (the late Zaha Hadid). Funky, illuminating facades take centre stage in these edifices in a bid to attract venture capitalists and encourage breakthroughs.

SOCIAL EXPERIMENT

The authors argue that this trend towards ‘luxe labs’ is a grand social experiment, with scientists as guinea pigs. They veer into an ethnographic study of researchers and their relationships to these buildings and breakout spaces, eavesdropping on their lunch conversations. They often cite the 1979 book *Laboratory Life* by ▶

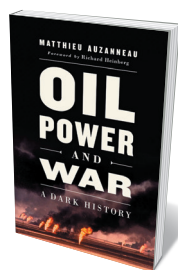
Books in brief



The Invisible Killer

Gary Fuller MELVILLE HOUSE (2018)

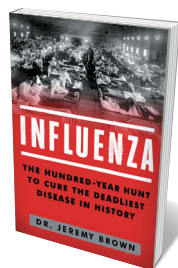
More than 90% of humanity is exposed to air-pollution concentrations exceeding World Health Organization guidelines. For this compelling exploration of an insidious crisis, air-quality researcher Gary Fuller travelled deep into our fume-ridden past. Here are seventeenth-century arborist John Evelyn’s observations of coal-burning in London; John Switzer Owens’s 1910s particulate gauges; longitudinal mortality research such as the 1993 US Six Cities study; impact analyses of lead fuels, diesel, biomass burning and land use; and a look at our current policy battle to breathe easy.



Oil, Power and War

Matthieu Auzanneau, transl. John F. Reynolds CHELSEA GREEN (2018)

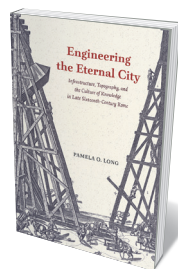
Oil is the dirty underlay to our times, reminds journalist Matthieu Auzanneau in this prodigious chronicle of the ‘fossil century’. Translated from French by John Reynolds, it is illuminating on the cascade of booms, busts, spills and quests for “nonconventional” sources such as shale. But Auzanneau extracts much more, showing how oil has shaped wars (for instance, through the decisive role of US fuel in British military aviation), Western and Arabic states, and dynasties such as the US Bush family, even as it foments environmental destruction. Auzanneau has created a towering telling of a dark and dangerous addiction.



Influenza

Jeremy Brown TOUCHSTONE (2018)

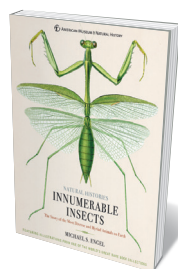
We should not underestimate influenza as a serial killer, notes physician Jeremy Brown in this agile study. Brown — director of emergency-care research at the US National Institutes of Health — illuminates much. Here is the science on viruses, those tiny replicating enigmas; outbreaks, from the catastrophic global 1918 Spanish flu pandemic to the 2002–03 SARS incident in which 10% of more than 8,000 people infected died; the complexities of data gathering, forecasting, drug stockpiling and vaccine hunting; and the lure of a cure. A thoughtful portrait of an elusive enemy.



Engineering the Eternal City

Pamela O. Long UNIVERSITY OF CHICAGO PRESS (2018)

For an ‘eternal’ city, Rome is hardly set in stone — and the late sixteenth century was one of its most fluid, architecturally. In this sparkling scholarly treatise, historian Pamela Long reveals how tottering infrastructure, ancient ruins and the flood-prone river Tiber were tamed by four successive popes with bold plans for the urban fabric. Drawing on a trove of archival maps and plans, Long charts the making and remaking of squares, aqueducts, sewers, streets and bridges — and engineer-hero Domenico Fontana’s stupendous feat in moving a 300-tonne obelisk to front St Peter’s Basilica.



Innumerable Insects

Michael S. Engel STERLING (2018)

Anyone who has thrilled to the shrilling of cicadas or marvelled at the bizarre behaviour of praying mantises will be entranced by this homage to the class Insecta. Distinguished entomologist Michael Engel has mined the library of New York’s American Museum of Natural History, and the spectacular images on show here — by Maria Sibylla Merian, John O. Westwood and many other greats of natural-history illustration — glow like jewels in a casket. With Engel’s deft text, this is a wonderful way to explore the riches of insect orders, from Blattodea to Zygentoma. **Barbara Kiser**



The Blizzard Building at Queen Mary University of London has sunken labs and elevated meeting pods.

► sociologists Bruno Latour and Steve Woolgar, who shadowed Salk staff like anthropologists, and argued that scientists' social interactions govern which lines of enquiry are ultimately pursued.

Scientists, however, are everyday humans. Do they need lavish surroundings or unusual furniture to trigger intellectual discussion?

The designers of the Blizzard Building, the biomedical hub of Queen Mary University of London, certainly thought so. Among its fantastical architectural elements are “mushroom”, “cloud” and “spikey” pods

serving as meeting and lounge spaces that “hover over the subterranean laboratories below”. The lab benches are standard, but sunken. I imagine researchers' annoyance at climbing stairs to get to their nearby desks, or wondering over the whimsical meeting spaces, when a few tables by the large windows would do.

WHERE'S THE EVIDENCE?

It is all very well for the physicists at the Perimeter Institute in Waterloo, Canada, to feel they can scribble on the vast windows, but spaces where experiments

happen must be practical and utilitarian. And contemplation can occur anywhere — in the shower, on a commute, while hiking (the Santa Fe Institute in New Mexico, set among hills and thermal pools, appreciates this). So far, no one has investigated whether Google engineers zipping around on Razor scooters to Lego-building stations innovate more freely than do their counterparts at more strait-laced firms.

In general, there seems to be a notable lack of consultation between architects and people who will work in their creations. One exception is the 2015

“There seems to be a notable lack of consultation between architects and people who will work in their creations.”

National Graphene Institute (NGI) on the campus of the University of Manchester, UK. Designers collaborated with institute researchers to yield a beautiful, functional building with easily adaptable clean rooms and other lab spaces enclosed by glass that invite both light and transparency around the work. Contributors Albena Yaneva and Stelios Zavos conclude that the NGI's labs actively shape and regulate the research culture, promoting “ecologies of innovation”, and “new alliances of science, society, and industry”.

But the authors' own photos show atria and plentiful couches devoid of humans (although they probably make nice napping platforms for overworked postdocs). Without evidence, it is an over-reach to say that the building's design accomplishes these grand goals.

I really wanted to see a controlled study on the nexus of built environment and research productivity. How difficult would it be to compare the output from researchers in the sleek NGI with that of those in an antiquated Manchester lab? Or to see whether Salk scientists in sunny La Jolla have made more breakthroughs than their counterparts in dreary basement labs at Mayo Clinic in Rochester, Minnesota?

Stranger even is the assertion at the end of *Laboratory Lifestyles*: that the dawn of the “petabyte age” of big data will make scientists and their hypotheses — and presumably labs — obsolete. Is this book an exploration of the lab or a prediction of its demise? In any case, it throws considerable doubt on whether some prominent lab architects understand the very passions that make lab occupants tick. ■

Kendall Powell is a freelance science journalist based in Lafayette, Colorado.
e-mail: kendallpowellsciwriting@gmail.com

Correspondence

Plan S to hit societies hard

Plan S is good news (*Nature* **561**, 17–18; 2018). As we move towards this subscription-free publishing model for 2020, the enormous costs that institutions pay to access the scientific literature will gradually be phased out. However, this could adversely affect the activities of academic societies that run their own journals.

These societies currently use income from subscription fees to host affordable conferences, run workshops, award travel grants, develop policy and engage in outreach. As publication charges for authors replace subscription fees, this income will plummet. Funding bodies might need to step in to make up the shortfall for supporting these services to the scientific community.

When publication charges become the norm, authors who cannot afford to pay them must not be unfairly disadvantaged (see J. Measey *Nature* **562**, 494; 2018). Means-tested rules for fee waivers will need to be factored in to the new publishing model. **Michael Jennions**, **Rob Lanfear** *Australian National University, Canberra, Australia.*

Shinichi Nakagawa *University of New South Wales, Sydney, Australia.*
s.nakagawa@unsw.edu.au

University voices in climate negotiations

Research institutions are appointed to act as official ‘observer’ delegates at international climate negotiations that are hosted by the United Nations and are otherwise closed to journalists and the outside world (see go.nature.com/2atycmq). As non-party stakeholders, they will provide a layer of transparency at this week’s 24th annual Conference of the Parties session, for example. Thanks to the University Climate Delegation Coalition (UCDC) that we launched last year, these

delegates are no longer simply observers: they can now bring a wide range of research voices to the table.

As knowledge producers, climate delegates from research institutions are in a position to provide insight into and attention to climate policy. The UCDC aims to engage delegates across US institutions on common initiatives. Over several months, researchers talk to their delegate representatives about their priorities for climate-related policy topics — for example, for emissions inventories, technology transfer, ecosystem management and human rights.

University delegations therefore provide an opportunity for the broader research community to connect with international climate negotiations and with climate advocacy. **Samantha Basile***, **Michael Lerner*** *University of Michigan, Ann Arbor, Michigan, USA.*

Keyon Rostamnezhad *Northeastern University, Boston, Massachusetts, USA.*

**Competing interests declared (see go.nature.com/2rcnrdb for details).*
sjbasile@umich.edu

Brazil politics threat to food security

Last month’s 14th meeting of the Conference of the Parties to the Convention on Biological Diversity discussed the pressing issue of biodiversity conservation and its relation to food security. Brazil was a participant with a right to vote, although it had acted merely as an observer in negotiations on the 2010 Nagoya Protocol, which it had failed to ratify because of its agribusiness and other interests. In our view, Jair Bolsonaro’s incoming government is likely to stand by those interests, despite the need to protect one of the world’s most biodiverse countries (see also *Nature* **563**, 5–6; 2018).

The government now taking shape is committed to relaxing requirements for

environmental licences and loosening environmental regulations. The newly appointed minister of agriculture has called for measures such as the ‘pesticide package’ (bill number 6.299/2002), which would weaken the criteria for pesticide approval — despite the concerns of United Nations rapporteurs Hilal Elver (on the right to food) and Baskut Tuncak (on toxins).

And, in a further blow to biodiversity, Bolsonaro has promised to open up the Amazon for agribusiness, with no indication that he intends to support traditional communities. **Marina Demaria Venâncio** *Federal University of Santa Catarina, Florianópolis, Brazil.*
Kamila Pope, **Stefan Sieber** *Leibniz-Centre for Agricultural Landscape Research, Müncheberg, Germany.*
popekamilla@gmail.com

Dynamic tolls are no easy traffic fix

Peter Cramton and colleagues suggest that dynamic road pricing could be a solution to traffic problems (*Nature* **560**, 23–25; 2018). As social scientists, we argue that getting drivers to change their behaviour might not be so simple, because the behaviour does not depend only on prices.

For instance, drivers tend to stick with their usual routes, departure times and destinations. Many would be reluctant to adapt their trips to a road-pricing scheme that fluctuates across time and place according to traffic conditions, because this requires too much mental effort. And charges would need to be prohibitively high to persuade them to give up the convenience, independence, flexibility, comfort and speed of using their cars.

Neither can public support for road pricing be assumed. It was blocked in Manchester (2005) and Edinburgh (2007) in the United Kingdom, in the Netherlands in 2010, and in Copenhagen in 2012. Pricing policies need to be seen

as fair to be acceptable, which is more likely if they protect the environment and future generations. Such psychological motives are rarely considered in economic models and in public and policy debates.

More interdisciplinary research into the causes of traffic problems is necessary for designing socially feasible policy solutions. For example, public support could be grown by communicating the extent to which such schemes would meet their objectives and how drivers would benefit.

Geertje Schuitema *University College Dublin, Ireland.*
Linda Steg *University of Groningen, the Netherlands.*
geertje.schuitema@ucd.ie

Helmholtz mentored many luminaries

In addition to his own discoveries (see H. Schmidgen *Nature* **561**, 175; 2018), polymath Hermann von Helmholtz influenced the development of a whole group of illustrious physicists.

In 1879, for example, he advised his student Heinrich Rudolf Hertz to experimentally test the assumptions underlying James Clerk Maxwell’s theory of electromagnetism. Hertz subsequently became the first to demonstrate the existence of electromagnetic waves, which eventually led to the radio and to telecommunications. Helmholtz’s students and research associates also included Max Planck, Heinrich Kayser, Eugen Goldstein, Wilhelm Wien, Arthur König, Henry Augustus Rowland, Albert Abraham Michelson and Michael Pupin, several of whom went on to receive the Nobel prize.

Helmholtz’s insights continue to be pertinent today (see, for example, S. A. Khan *Int. J. Light Electron Opt.* **127**, 9798–9809; 2016).

Sameen Ahmed Khan *Dhofar University, Salalah, Oman.*
rohelaakhan@yahoo.com

Stiff competition

ARISING FROM J. B. Berger, H. N. G. Wadley & R. M. McMeeking, *Nature* **543**, 533–537 (2017); <https://doi.org/10.1038/nature21075>

The paper of Berger, Wadley & McMeeking¹ presents beautiful results on structured composites near the edge of maximal stiffness for a given porosity. However, it appears that the authors were unaware of the large body of work on this subject, much of it summarized in refs 2–6. In particular, their claim that “a material geometry that achieves the theoretical upper bounds for isotropic elasticity and strain energy storage (the Hashin–Shtrikman upper bounds) has yet to be identified” is not accurate. Multiscale elastically isotropic composites with simultaneously maximal bulk and shear modulus—and hence maximal stiffness and energy storage—were identified independently in refs 7–9. There is a Reply to this Comment by J. B. Berger et al., *Nature* **564**, <https://doi.org/10.1038/s41586-018-0725-7> (2018).

Moreover, the simple argument made in ref. 8—that the Hashin–Shtrikman bounds are attained if the actual field in the material matches the trial field, which is constant in one phase—shows that any hierarchical laminate, in which layers of the stiffer phase are sequentially added to the composite in different orientations, necessarily achieves these upper bounds if layering is done so that the final material is elastically isotropic. Later it was established by Bourdin and Kohn¹⁰ that no separation of length scales is needed if the volume fraction of the stiffer phase is small. These geometries are formed by the union of families of parallel plates, with each family having a different orientation, and include the cubic foam, octet foam and cubic + octet foam described in ref. 1. The novelty of ref. 1 is that it shows that this class of microstructure also works well if the volume fraction is moderate. Other porous three-dimensional microgeometries with very large bulk and shear moduli, at a moderate volume fraction of 0.338, have been found using topology optimization methods¹¹ (see, in particular, point e in figure 9 of ref. 11). Yet it is still not known if a single-scale geometry can exactly attain the shear bounds away from the low-density limit. Single-scale geometries can achieve the bulk modulus bounds^{12–14}.

To finish, we briefly mention important results that cover more general questions than those addressed in ref. 1 to bring readers up to speed on current developments. If the second material is not void, there are improved bounds that couple the possible bulk and shear moduli^{15,16}, and the range of possible (bulk, shear) pairs has been explored numerically^{11,17}. A recent paper¹⁸ goes a long way to completely characterizing the possible elasticity tensors of three-dimensional printed, possibly anisotropic, materials constructed from a given isotropic material with given porosity. These materials include elastically isotropic microstructures that asymptotically attain the Hashin–Shtrikman upper bulk modulus bound for any given volume fraction, yet have an arbitrarily small shear modulus. If one allows the starting material to be as stiff as one likes, and replaces the void material by a material that is as compliant as one likes, then one can get any desired elasticity tensor¹⁹—a result also suggested by numerics¹⁷. In fact, non-local effective behaviours are possible too and, remarkably, these have also been completely characterized for linear elasticity²⁰. In principle, one can obtain composites for which uniform strains cost little energy, but gradients in the strains (double gradients of the displacement) cost considerable energy (see, for example, ref. 21 for some interesting examples).

Data availability

All data are available from the corresponding author upon reasonable request.

G. W. Milton^{1*}

¹Department of Mathematics, The University of Utah, Salt Lake City, UT, USA. *e-mail: milton@math.utah.edu

Received: 2 February; Accepted: 12 September 2018;

Published online 5 December 2018.

- Berger, J. B., Wadley, H. N. G. & McMeeking, R. M. Mechanical metamaterials at the theoretical limit of isotropic elastic stiffness. *Nature* **543**, 533–537 (2017).
- Cherkaev, A. *Variational Methods for Structural Optimization* (Springer-Verlag, New York, 2000).
- Milton, G. W. *The Theory of Composites* (Cambridge Univ. Press, Cambridge, 2002).
- Allaire, G. *Shape Optimization by the Homogenization Method* (Springer-Verlag, New York, 2012).
- Torquato, S. *Random Heterogeneous Materials: Microstructure and Macroscopic Properties* (Springer Science & Business Media, New York, 2002).
- Tartar, L. *The General Theory of Homogenization: A Personalized Introduction*, (Springer-Verlag, Berlin, Heidelberg, 2009).
- Norris, A. N. A differential scheme for the effective moduli of composites. *Mech. Mater.* **4**, 1–16 (1985).
- Milton, G. W. in *Homogenization and Effective Moduli of Materials and Media* (eds Ericksen, J. L. et al.) 150–174 (Springer-Verlag, New York, 1986).
- Francfort, G. A. & Murat, F. Homogenization and optimal bounds in linear elasticity. *Arch. Ration. Mech. Anal.* **94**, 307–334 (1986).
- Bourdin, B. & Kohn, R. V. Optimization of structural topology in the high-porosity regime. *J. Mech. Phys. Solids* **56**, 1043–1064 (2008).
- Andreassen, E., Lazarov, B. S. & Sigmund, O. Design of manufacturable 3D extremal elastic microstructure. *Mech. Mater.* **69**, 1–10 (2014).
- Vigdergauz, S. B. Effective elastic parameters of a plate with a regular system of equal-strength holes (Effektivnye uprugie parametry plastiny s reguliarnoi sistemoi ravnoprochnykh otverstii). *Inzh. Zh. Mekh. Tver. Tela.* **21**, 165–169 (1986).
- Grabovsky, Y. & Kohn, R. V. Microstructures minimizing the energy of a two phase elastic composite in two space dimensions. II. *The Vigdergauz microstructure*. *J. Mech. Phys. Solids* **43**, 949–972 (1995).
- Liu, L., James, R. D. & Leo, P. H. Periodic inclusion-matrix microstructures with constant field inclusions. *Metall. Mater. Trans. A* **38**, 781–787 (2007).
- Berryman, J. G. & Milton, G. W. Microgeometry of random composites and porous media. *J. Phys. D* **21**, 87–94 (1988).
- Cherkaev, A. V. & Gibiansky, L. V. Coupled estimates for the bulk and shear moduli of a two-dimensional isotropic elastic composite. *J. Mech. Phys. Solids* **41**, 937–980 (1993).
- Sigmund, O. Materials with prescribed constitutive parameters: an inverse homogenization problem. *Int. J. Solids Struct.* **31**, 2313–2329 (1994).
- Milton, G. W., Briane, M. & Harutyunyan, D. On the possible effective elasticity tensors of 2-dimensional and 3-dimensional printed materials. *Math. Mech. Complex Syst.* **5**, 41–94 (2017).
- Milton, G. W. & Cherkaev, A. V. Which elasticity tensors are realizable? *J. Eng. Mater. Technol.* **117**, 483–493 (1995).
- Camar-Eddine, M. & Seppecher, P. Determination of the closure of the set of elasticity functionals. *Arch. Ration. Mech. Anal.* **170**, 211–245 (2003).
- Seppecher, P., Alibert, J.-J. & dell’Isola, F. Linear elastic trusses leading to continua with exotic mechanical interactions. *J. Phys.* **319**, 012018 (2011).

Competing interests Declared none.

Additional information

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to G.W.M.

<https://doi.org/10.1038/s41586-018-0724-8>

Berger et al. reply

REPLYING TO G. W. Milton, *Nature* **564**, <https://doi.org/10.1038/s41586-018-0724-8> (2018)

In the accompanying Comment¹, Milton correctly points out that material geometries that achieve the Hashin–Shtrikman upper bounds² have been previously identified. We thank G. W. Milton for his interest in our work³ and welcome his insights, including those that look beyond our contribution in regard to elasticities that are realizable.

Milton correctly refutes our statement that “a material geometry that achieves the theoretical upper bounds for isotropic elasticity and strain energy storage (the Hashin–Shtrikman upper bounds²) has yet to be identified”. In his Comment¹, Milton has pointed to many studies that have identified material combinations and multi-length-scale geometries that achieve the Hashin–Shtrikman² theoretical upper bound^{4–6}. We acknowledge and accept this correction of our claim.

Retrospectively, it is clear that we should have qualified our claim and placed it in the narrower context of our study (namely, the design of a single-length-scale, single-material, elastically isotropic lattice that is easily fabricated)—a context motivated by the need for lightweighting and the continuing discovery of multifunctional structural systems. We did describe (to our knowledge, for the first time) a single-length-scale biphasic material geometry—specifically, a combination of void and solid phases—that performs at, or nearly at, the Hashin–Shtrikman upper bounds for both the bulk and shear moduli simultaneously, over a wide range of relative lattice densities. This design is simple and manufacturable and was demonstrated to achieve, or nearly achieve, the Hashin–Shtrikman² theoretical upper bounds. In addition, because it is composed of two anisotropic but maximally efficient sub-geometries, it enables the creation of multifunctional lightweight structures. In our paper³, we provided analytical proof of our design’s maximal elastic performance, as well as numerical evidence of its optimal elastic performance over a wide range of relative densities. In this restricted ‘single-length-scale, single-material’ context, we assess our claim to be accurate, and accept that we were remiss in not stating this context more clearly.

The summarizing works^{7–10} cited by Milton present techniques for generating optimal material microstructures. However, none directly addresses the problem that we sought to solve. Milton states that

material geometries that achieve the Hashin–Shtrikman upper bounds simultaneously have previously been identified^{4–6}. Although this is true, we find that there are notable differences that clearly differentiate our work from these studies. These are perhaps most evident in the geometric simplicity of our design and its implications for the fabricability, and therefore the utility, of our design as an engineering material system.

Although Norris⁴ identified a microstructure that simultaneously achieves the Hashin–Shtrikman upper bounds, this solution consists of solid disks embedded in vacuum, which is impractical. Francfort and Murat⁶ proved mathematically that laminates that stack in three dimensions can also simultaneously achieve the Hashin–Shtrikman upper bounds. However, the authors specify that both phases are solid, so low-density, single-solid-material systems with void space, such as ours, are not accessible. In both cases, voided regions can be approximated as a very-low-density phase. However, the design of this porous phase is still an issue, which is essentially identical to the fundamental problem of identifying a single-length-scale maximally stiff isotropic material geometry. This only adds to the difficulty of the solution by requiring material geometries to be constructed at even smaller length scales.

Ranked laminates have previously been shown to simultaneously achieve the Hashin–Shtrikman upper bounds⁵. Such laminates rely on multiple length scales, and at each level the smaller-scale composites are assumed to be isotropic and effectively continuous. We purposely avoid such complexity in our approach, in an effort to achieve simple and therefore manufacturable geometries.

Bourdin and Kohn¹¹ studied a family of material geometries that contain parallel planes of material and that would appear to contain the cubic and octet geometries. These materials were found to simultaneously achieve the Hashin–Shtrikman upper bounds in the low-density limit—an aspect paralleled in our work. The authors performed numerical calculations to obtain two-dimensional solutions but did not go on to generate three-dimensional designs—thus avoiding what might be the most practical application of ref. ¹¹. Our work does parallel ref. ¹¹ in the recognition that sheets of material are required for optimal performance. This insight, regrettably, appears to have been overlooked

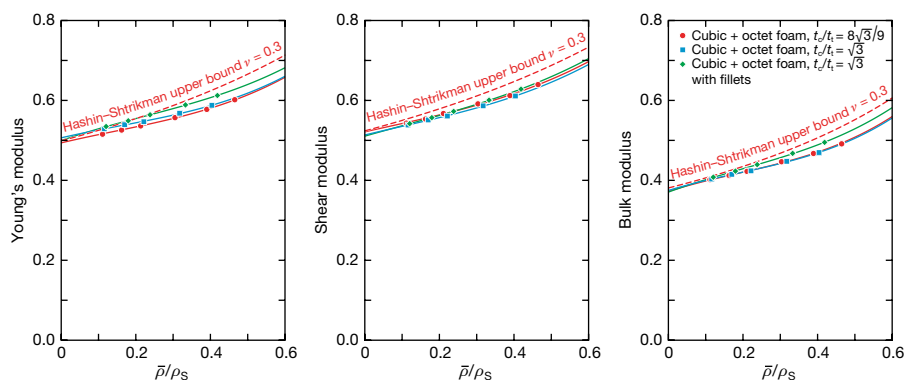


Fig. 1 | Young's, shear and bulk moduli. Finite-element analysis results indicate that with the addition of fillets, the normalized elastic moduli of the cubic + octet foam can achieve more than 98% of the theoretical upper limit for specific strain energy storage when the relative density is moderate to low, $\bar{\rho}/\rho_s \leq 26.4\%$; $\bar{\rho}$ is the effective density of the cellular material and ρ_s is the density of the constituent (solid) material (ν is the

Poisson ratio). The filleted design with a wall thickness ratio of $t_c/t_t = \sqrt{3}$ at $\bar{\rho}/\rho_s = 26.4\%$ has a shear performance that reaches 96.7% of the Hashin–Shtrikman upper limit. By varying the wall thickness ratio, isotropy can be achieved independently of the total performance, so that 98% of the shear upper bound can be realized at this relative density.

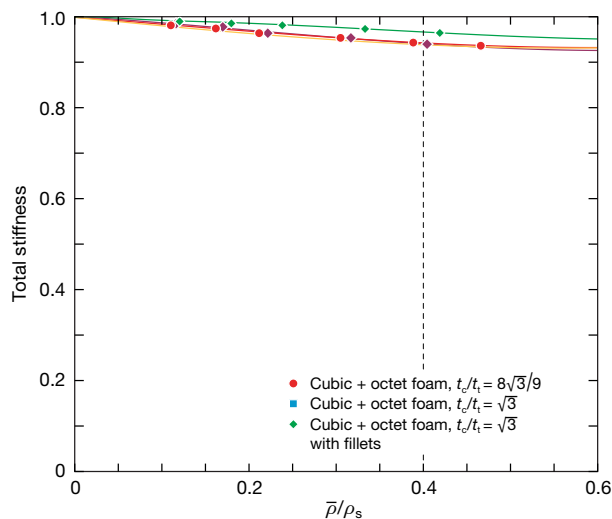


Fig. 2 | Total stiffness. The total specific strain energy storage of the cubic + octet foam is enhanced by the addition of fillets. This topology achieves more than 96.6% of the upper bounds when $\bar{\rho}/\rho_s \leq 40.0\%$.

by the broader community, and we are unaware of any subsequent study that utilized it as a design principle to identify a manufacturable three-dimensional solution such as ours.

There are numerous studies that involve mathematical approaches to achieve extremal performance in composite systems. One of the primary issues with the further development of these systems into structural materials is the often-complex nature of these designs. For example, Cherkashev⁷ describes composites that are assumed to be composed of small fragments, which are necessarily much smaller than the domain under consideration, with ranked laminates being a subset of these materials. The practicality of fabricating such complex geometries is questionable, since Cherkashev himself suggests that these are most useful as design guidelines for more practical approaches⁷.

There is certainly a large and interesting body of work in the area of topology and geometry optimization that address a space that encompasses and goes beyond the scope of our work—some of which Milton highlights^{12–18}. We appreciate his identification of the parallels between our work and this important area of study.

While it is still not known whether a material geometry exists that can achieve the Hashin–Shtrikman shear upper bound away from the low-density limit, the cubic + octet foam achieves 94.7% of the Hashin–Shtrikman upper bound on shear modulus, and 95.2% of the bound on Young’s modulus, at a moderate relative density of 0.338, while having a Zener anisotropy ratio of 1.01. By reducing stress concentrations by rounding the joints where webs intersect (that is, with the addition of fillets), these can be increased to 96.4%, 98.1% and 1.02, respectively (this is with a wall thickness ratio of $t_c/t_t = \sqrt{3}$ —not $t_c/t_t = 8\sqrt{3}/9$, which is isotropic in the low-density limit³; t_c and t_t are the wall thicknesses of the cubic and octet sub-geometries, respectively) (Fig. 1). These can easily be made isotropic by varying the ratio of the wall thicknesses, t_c/t_t . This improvement is not the result of rigorous optimization, but rather a simple ad hoc approach that leaves room for potentially even better-performing designs. If the results of Andreassen et al.¹⁹ are indeed relevant and noteworthy, as Milton points out, then it pays to mention that the cubic + octet foam does, in essence, achieve the theoretical upper bounds for structural efficiency away from the low-density limit (Fig. 2), including that for shear modulus, and that the identification of such a material geometry is not a completely open problem.

In the papers discussed above we find no description or illustration of a simple three-dimensional, low-density geometry with a single length scale and fabricated using a single solid material, that achieves the Hashin–Shtrikman upper bounds on elastic moduli; that is, we find nothing similar to the geometry that we have developed. Although the work of Bourdin and Kohn¹¹ does appear to facilitate the generation of our extremal design, the authors do not use their numerical scheme to solve any three-dimensional problems, and their proofs address only the low-density limit. We acknowledge that Professor Milton has provided a helpful summary of theoretical approaches that complement the approach that we have taken. However, considering the limited space available for the presentation of our study, the focus of our paper³ was to describe and discuss our design approach and the mechanical properties of the resulting topology. This focus determined the emphasis of the work that we presented and restricted the literature that we selected to cite.

Data availability

The datasets generated and analysed during the current study are available from the corresponding author on reasonable request.

J. B. Berger^{1,2*}, H. N. G. Wadley³ & R. M. McMeeking^{1,2,4,5}

¹Materials Department, University of California, Santa Barbara, CA, USA.

²Department of Mechanical Engineering, University of California, Santa

Barbara, CA, USA. ³Department of Materials Science and Engineering,

School of Engineering and Applied Science, University of Virginia,

Charlottesville, VA, USA. ⁴School of Engineering, University of Aberdeen,

King’s College, Aberdeen, UK. ⁵INM-Leibniz Institute for New Materials,

Saarbrücken, Germany. *e-mail: berger@engineering.ucsb.edu

Published online 5 December 2018.

1. Milton, G. W., *Nature* **564**, <https://doi.org/10.1038/s41586-018-0724-8> (2018).
2. Hashin, Z. & Shtrikman, S. A variational approach to the theory of the elastic behaviour of multiphase materials. *J. Mech. Phys. Solids* **11**, 127–140 (1963).
3. Berger, J. B., Wadley, H. N. G. & McMeeking, R. M. Mechanical metamaterials at the theoretical limit of isotropic elastic stiffness. *Nature* **543**, 533–537 (2017).
4. Norris, A. N. A differential scheme for the effective moduli of composites. *Mech. Mater.* **4**, 1–16 (1985).
5. Milton, G. W. in *Homogenization and Effective Moduli of Materials and Media* (eds Ericksen, J. L. et al.) 150–174 (Springer-Verlag, New York, 1986).
6. Francfort, G. A. & Murat, F. Homogenization and optimal bounds in linear elasticity. *Arch. Ration. Mech. Anal.* **94**, 307–334 (1986).
7. Cherkashev, A. *Variational Methods for Structural Optimization* (Springer-Verlag, New York, 2000).
8. Milton, G. W. *The Theory of Composites* (Cambridge Univ. Press, Cambridge, 2002).
9. Allaire, G. *Shape Optimization by the Homogenization Method* (Springer-Verlag, New York, 2012).
10. Torquato, S. *Random Heterogeneous Materials: Microstructure and Macroscopic Properties* (Springer Science & Business Media, New York, 2002).
11. Bourdin, B. & Kohn, R. V. Optimization of structural topology in the high-porosity regime. *J. Mech. Phys. Solids* **56**, 1043–1064 (2008).
12. Berryman, J. G. & Milton, G. W. Microgeometry of random composites and porous media. *J. Phys. D* **21**, 87–94 (1988).
13. Cherkashev, A. V. & Gibiansky, L. V. Coupled estimates for the bulk and shear moduli of a two-dimensional isotropic elastic composite. *J. Mech. Phys. Solids* **41**, 937–980 (1993).
14. Sigmund, O. Materials with prescribed constitutive parameters: an inverse homogenization problem. *Int. J. Solids Struct.* **31**, 2313–2329 (1994).
15. Milton, G. W. & Cherkashev, A. V. Which elasticity tensors are realizable? *J. Eng. Mater. Technol.* **117**, 483–493 (1995).
16. Milton, G. W., Briane, M. & Harutyunyan, D. On the possible effective elasticity tensors of 2-dimensional and 3-dimensional printed materials. *Math. Mech. Complex Syst.* **5**, 41–94 (2017).
17. Camar-Eddine, M. & Seppecher, P. Determination of the closure of the set of elasticity functionals. *Arch. Ration. Mech. Anal.* **170**, 211–245 (2003).
18. Seppecher, P., Alibert, J.-J. & dell’Isola, F. Linear elastic trusses leading to continua with exotic mechanical interactions. *J. Phys.* **319**, 012018 (2011).

BRIEF COMMUNICATIONS ARISING

19. Andreassen, E., Lazarov, B. S. & Sigmund, O. Design of manufacturable 3D extremal elastic microstructure. *Mech. Mater.* **69**, 1–10 (2014).

Author contributions J.B.B. created the ideas, conceived and designed the new material geometries and performed the structural analysis. R.M.M. developed the analytical models for the strain energy and moduli and, with H.N.G.W., contributed to refining the concepts, contextualizing the results and providing critiques and assessments.

Competing interests The material geometry identified in this work to achieve the theoretical bounds in performance has been included in a Patent

Cooperation Treaty (PCT/US2015/010458) by Nama Development, LLC (DE), which is majority-owned by J.B.B.

Additional information

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to J.B.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

<https://doi.org/10.1038/s41586-018-0725-7>

GEOCHEMISTRY

The rocky road to biomolecules

A natural chemical reaction that occurs below the sea floor makes the amino acid tryptophan without biological input. This finding reveals a process that might have helped life on Earth to begin. [SEE ARTICLE P.59](#)

JOHN A. BAROSS

Robert Frost's poem *Fire and Ice* ponders which of these two will eventually cause life on Earth to cease. Conversely, scientists have long been interested in whether life on this planet originated in hot or cold conditions. Did life arise in a hot volcanic environment^{1,2} or, as Charles Darwin suggested in a letter (see go.nature.com/2q8w3n5), in "some warm little pond"? Might we need to invoke a global setting that includes ice³ or would an ocean-floor setting suffice^{1–5}? On page 59, Ménez *et al.*⁶ report an analysis of rocky material below the ocean floor at sites called serpentinizing hydrothermal vents, which are hot springs that discharge alkaline, gas-rich water. The authors provide evidence that a chemical reaction occurs there that could have set the stage for life to begin — the generation of an amino acid by a process that is not biologically mediated.

A well-characterized form of hydrothermal vent on the ocean floor, called black-smoker chimneys, occurs in a volcanic, magma-rich setting. These vents emit acidic fluids into the ocean that have high concentrations of gases and also of metals that are in a chemically reduced form.

However, in 2000, the serendipitous discovery of a mid-Atlantic ocean-floor site called Lost City revealed another type of hydrothermal-vent environment⁷. The Lost City vents were found to arise by a process termed serpentinization — a chemical interaction between water and a type of rock called peridotite that contains minerals enriched in magnesium, iron and silica. Serpentinization generates alkaline conditions that aid the formation of majestic carbonate-rich towers (Fig. 1). Serpentinization also produces hydrogen and a variety of organic molecules, including formate, acetate and pyruvate, which might be important in supporting the microbial life at Lost City and could also have been used in the biochemical steps that led to life on Earth^{8,9}. Studies of Lost City have provided a treasure trove of scientific discoveries that have greatly altered our understanding of hydrothermal vents and the geological history of the early Earth¹⁰. They have also led to testable hypotheses about possible

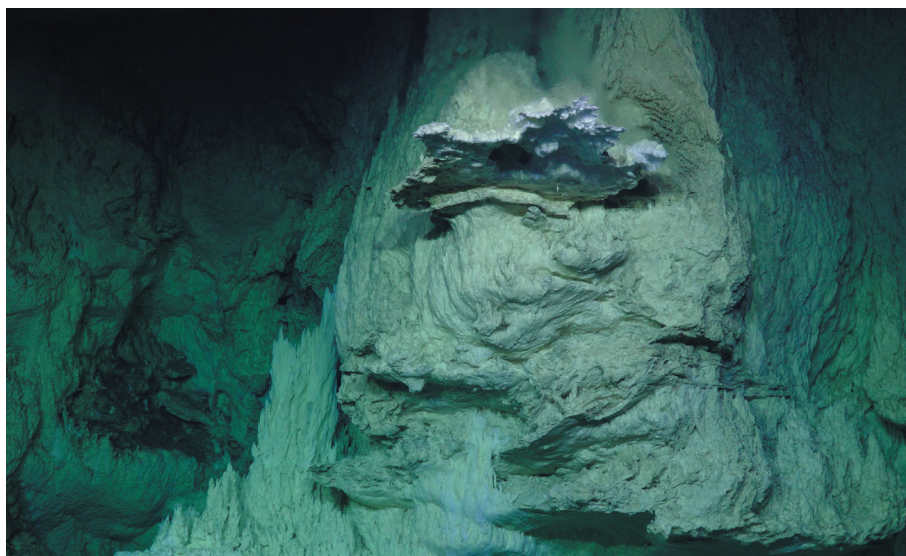


Figure 1 | Carbonate towers in hydrothermal vents at the Lost City site on the floor of the Atlantic Ocean.

environmental settings for the origin of life.

In laboratory experiments that simulate conditions in magma-hosted hydrothermal vents, amino acids can be synthesized by chemical reactions¹¹ that do not require biological input. However, whether such abiotic generation of amino acids can occur in serpentinizing hydrothermal vents was unknown. Amino acids have been detected¹² in fluids emanating from Lost City, but their source was undetermined.

To investigate further, Ménez and colleagues analysed material from rock samples retrieved by drilling around 170 metres below the ocean floor at Lost City. The authors present extensive evidence for the presence of the nitrogen-containing amino acid tryptophan in a context in which it was unlikely to have been produced by a biologically mediated process. They report data obtained from three high-resolution techniques that are consistent with the presence of tryptophan. The researchers also found other organic compounds that might be intermediates in the synthesis of tryptophan. Ménez *et al.* propose that this synthesis could be the result of Friedel–Crafts reactions because they found the molecule indole, which is an intermediate organic compound in the synthesis of tryptophan by this type of reaction. Their case for the abiotic synthesis of tryptophan is strengthened

by the absence of other amino acids that would be present if a biological source was there, such as microbial contamination of the rock sample.

To extend Ménez and co-workers' report of the abiotic synthesis of tryptophan, future studies at Lost City should try to collect adequate volumes of fluid to determine a structural property, called chirality, of the tryptophan present. Molecules can exist in two mirror-image chiral forms. Synthesis of a molecule by a non-biological process generally results in equal proportions of these two forms, whereas biologically synthesized amino acids are usually made in predominantly one form or the other.

The authors' work also sheds light on the long-standing mystery of what mechanism reduces nitrogen molecules (N_2) to ammonia under hydrothermal-vent conditions. In most of the cases in which ammonia has been detected in hydrothermal-vent environments, it was found to originate from buried sediment sources of organic material and not from abiotic synthesis in the hydrothermal vents¹³. Ménez and colleagues propose that saponite, an iron-containing clay mineral that they detected, and which is reported¹⁴ to be a catalyst for the synthesis of organic compounds and the reduction of N_2 to ammonia, might be involved in tryptophan synthesis. Abiotic generation of a source of ammonia, together

SUSAN LANG, UNIV. SOUTH CAROLINA/NSF/ROV JASON/2018. COPYRIGHT WOODS HOLE OCEANOGRAPHIC INSTITUTION

with saponite's proposed catalysis of heterocyclic-amine molecules such as tryptophan, also raises the possibility of abiotic synthesis of other heterocyclic amines called pyrimidines and purines, which are components of the nucleic acids DNA and RNA. Moreover, saponite has the potential to promote the formation of organic polymers^{15,16}.

Beyond the potential for the synthesis and accumulation of organic compounds that were probably important in the origin of life, serpentinization has two other characteristics that have intriguing implications regarding the origin of life and the establishment of habitable conditions^{5,8}. One characteristic is that serpentinization produces heat. The gradient of temperature can reach more than 200 °C at the site of the serpentinization reaction⁴. This, in turn, promotes hydration and therefore expansion of rocks, which is the other intriguing characteristic of serpentinization. However, if part of the ocean floor 'sinks' (subducts) into Earth's interior as tectonic plates move, the greater heat and pressure encountered on its descent into the deep subsurface region would reverse such serpentinization, and the water released during this reversal could help to give rise to volcanoes on the ocean floor¹⁴. This, in turn, might help to recycle key elements that support life.

The geological record at the time of life's origin, 3.5 billion to 4.4 billion years ago, is enriched in iron- and magnesium-containing minerals (characteristic of the rocks that form Earth's mafic crust), and in other elements that could have been extracted from rock in a process mediated by high-temperature water, strongly pointing to hydrothermal activity at that time. Yet, during the first billion years of Earth's history, the heat from the mantle was too great for plate tectonics to occur^{17,18}. Consequently, heat would have been lost from Earth's interior mainly through volcanoes on the ocean floor. Earth's crust would have been rich in silicate minerals and iron^{17,18}, allowing high rates of serpentinization and producing high concentrations of hydrogen and organic compounds.

Extensive circulation of seawater through volcanic rock during this time might have resulted in heat, fluid and gas ascending from the depths to create convective cells — a phenomenon characterized by currents due to density differences in the liquids or gases present. In the volcanic-rock environment, this could have led to associated gradients of temperature, pressure, chemical composition and wet-dry cycles (hydration-dehydration cycles known to promote chemical reactions that include the polymerization of organic compounds). Regardless of how life originated and in what environmental setting it was first established, serpentinization probably had an important role in facilitating the availability of organic chemicals required for life.

Understanding serpentinization at Lost City has wider scientific implications. Saturn's

icy moon Enceladus has many of the chemical properties known to support life^{19,20} that are seen in serpentinizing environments such as those of Lost City. Whether or not Enceladus, or indeed other icy moons such as Jupiter's Europa, could or did support life, they nevertheless could provide insight into geochemical processes that might lead to life. Such geochemical analysis seems to support the hypothesis that hydrothermal systems might have had an essential role in the origin of life. A more far-reaching implication of the work by Ménez and colleagues, and of others investigating hydrothermal vents, is that efforts to understand the characteristics of these settings might aid efforts to search for life beyond Earth. A planetary body with evidence of geophysical properties, including plate tectonics and hydrothermal systems, might have a higher probability of acquiring and supporting carbon-based life than planetary bodies lacking such geophysical properties. If true, then targeting such planets might also increase our probability of finding such life. ■

John A. Baross is in the School of Oceanography and the Astrobiology Program, University of Washington, Seattle, Washington 98195, USA.

e-mail: jbaross@u.washington.edu

1. Baross, J. A. & Hoffman, S. E. *Orig. Life Evol. Biosph.* **15**, 327–345 (1985).
2. Martin, W. & Russell, M. J. *Phil. Trans. R. Soc. Lond. B* **362**, 1887–1925 (2007).
3. Stüeken, E. E. *et al. Geobiology* **11**, 101–136 (2013).
4. Martin, W., Baross, J., Kelley, D. & Russell, M. J. *Nature Rev. Microbiol.* **6**, 805–814 (2008).
5. Preiner, M. *et al. Life* **8**, 41 (2018).
6. Ménez, B. *et al. Nature* **564**, 59–63 (2018).
7. Kelley, D. *et al. Nature* **412**, 145–149 (2001).
8. Lang, S. Q., Butterfield, D. A., Schulte, M., Kelley, D. S. & Lilley, M. D. *Geochim. Cosmochim. Acta* **74**, 941–952 (2010).
9. Schrenk, M. O., Brazelton, W. J. & Lang, S. Q. *Rev. Miner. Geochem.* **75**, 575–606 (2013).
10. Sleep, N. H., Bird, D. K. & Pope, E. C. *Phil. Trans. R. Soc. B* **366**, 2857–2869 (2011).
11. Hennet, R. J., Holm, N. G. & Engel, M. H. *Naturwissenschaften* **79**, 361–365 (1992).
12. Lang, S. Q., Früh-Green, G. L., Bernicconi, S. M. & Butterfield, D. A. *Geobiology* **11**, 154–169 (2013).
13. Lilley, M. D. *et al. Nature* **364**, 45–47 (1993).
14. Koobi, F. & Jones, W. *Clay Miner.* **32**, 633–643 (1997).
15. Ferris, J. P., Hill, A. R. Jr, Liu, R. & Orgel, L. E. *Nature* **381**, 59–61 (1996).
16. Hazen, R. M. & Sverjensky, D. A. *Cold Spring Harb. Perspect. Biol.* **2**, a002162 (2010).
17. Dhuime, B., Wuestefeld, A. & Hawkesworth, C. J. *Nature Geosci.* **8**, 552–555 (2015).
18. Tang, M., Chen, K. & Rudnick, R. L. *Science* **351**, 372–375 (2016).
19. Waite, J. H. *et al. Science* **311**, 1419–1422 (2006).
20. Waite, J. H. *et al. Science* **356**, 155–159 (2017).

This article was published online on 7 November 2018.

CONDENSED-MATTER PHYSICS

Elusive spin textures discovered

Magnetic materials can host a range of structures called spin textures. Two such textures — a meron and an antimeron — have been observed experimentally for the first time, in a material known as a chiral magnet. SEE LETTER P.95

SEONGHOON WOO

Magnetic moments (spins) in magnetic materials can form various structures known as spin textures. In most cases, spins of neighbouring atoms tend to align parallel or antiparallel to each other, resulting in ferromagnets or antiferromagnets, respectively. However, in some materials called chiral magnets that have unusual physical interactions between spins owing to a peculiar crystalline or multilayer structure, spins align in an intricate fashion: a topological spin texture. On page 95, Yu *et al.*¹ report the first experimental evidence for two such textures, and observe transitions between textures that could have applications in spin-based electronics (spintronics).

The archetypal topological spin texture is a small, swirling magnetic knot known as a magnetic skyrmion (Fig. 1). In a skyrmion, the

orientation of spins rotates progressively from the up direction at the edge of the texture to the down direction at the centre, or vice versa. The properties of a skyrmion can be characterized by a value of either -1 or $+1$ for a quantity called the topological charge.

Magnetic skyrmions were discovered² in 2009 and were observed at room temperature³ in 2015. Since then, they have been at the centre of research in many scientific and technical fields, for at least three reasons⁴. First, they can be very stable, owing to a phenomenon called topological protection. Second, they can be extremely tiny (with diameters in the nanometre range), which means that they could be used in future nanotechnology. And third, they exhibit energy-efficient current-driven behaviour that is suitable for next-generation low-energy spintronic devices such as those involving computer memory, logic, information transmission and

neuromorphic (brain-like) computing.

The remarkable properties of magnetic skyrmions have inspired strong worldwide research efforts looking for the other types of topological spin texture that could exist in chiral magnets: antiskyrmions, merons and antimerons (Fig. 1). In 2017, antiskyrmions were observed⁵ in chiral magnets called acentric tetragonal magnetic Heusler compounds, which have unusual crystal structures. However, merons and antimerons have been elusive.

Skyrmions were initially considered in many areas of physics other than the study of magnetic materials. Likewise, the concept of merons and antimerons originated in classical field theory⁶ and was later applied to particle physics⁷ and to condensed-matter systems such as quantum Hall materials⁸ and chiral magnets and spintronics⁹. In chiral magnets, merons and antimerons have different topological properties from skyrmions. In merons and antimerons, the spins at the core region point in the up or down direction, but those at the periphery align in the plane of the material, corresponding to a topological charge of $-\frac{1}{2}$ or $+\frac{1}{2}$.

Many theoretical studies in the past few years have asserted the possibility of the existence of merons and antimerons in chiral magnets that exhibit in-plane magnetic anisotropy^{10,11} — a property in which the response of the material to a magnetic field is largest when the field is directed along the plane of the material. Yu and colleagues tested this prediction using a thin film of a chiral magnet containing cobalt, zinc and manganese that has a cubic crystalline structure and in-plane magnetic anisotropy. The authors carefully engineered the material composition of the film, because they found that the composition had a substantial effect on spin texture.

Yu *et al.* then carried out highly sophisticated spin-visualization measurements of the synthesized magnetic film using a technique called Lorentz transmission electron microscopy. This technique could resolve spin orientations using the interaction between electrons from the microscope and the spins in the presence of a magnetic field. The authors performed spin-texture imaging at various temperatures and magnetic-field strengths. On the basis of these observations, the work presents several key achievements.

Yu and colleagues are the first to have observed merons and antimerons in a chiral magnet. When the authors applied a magnetic field of 20 millitesla perpendicular to the magnetic film at a temperature of 295 kelvin, they observed a stabilized square lattice of merons and antimerons. Such a lattice is characterized by periodic arrays of alternating convergence and divergence of magnetization. Unlike the typical hexagonal lattice of skyrmions, in which no antiskyrmions exist, it is noteworthy that the observed square lattice contained both merons and antimerons — a configuration that was often considered as a separate spin texture,

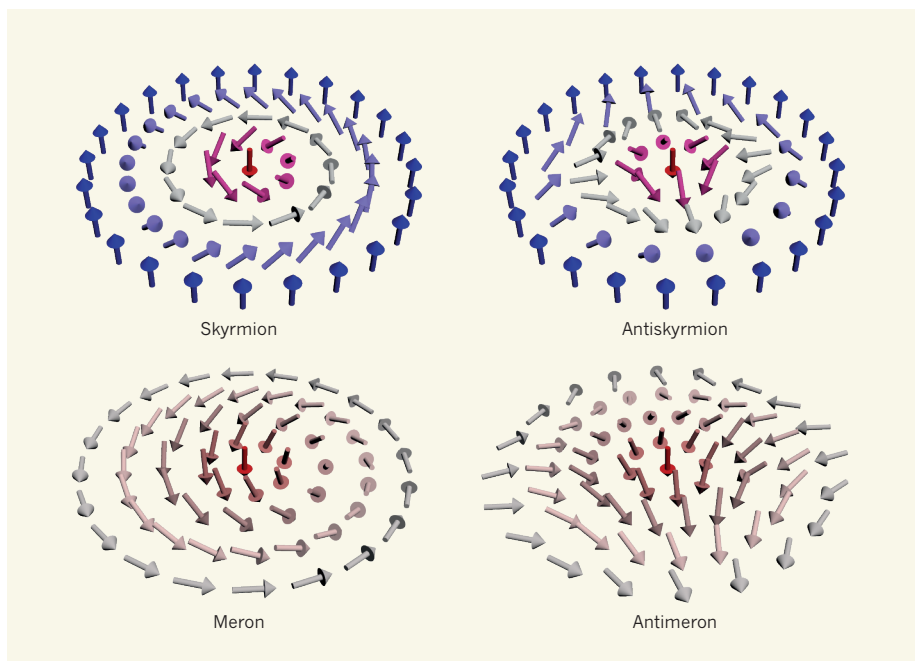


Figure 1 | Topological spin textures in a chiral magnet. In materials known as chiral magnets, magnetic moments (arrows) can form intricate patterns called topological spin textures. The four types of texture that could exist in a chiral magnet are skyrmions, antiskyrmions, merons and antimerons. Yu *et al.*¹ report the detection of merons and antimerons in a particular chiral magnet, and the observation of transformations between topological spin textures. The colours indicate the spatial direction of the magnetic moments out of the plane of the material, from the up direction (blue) to the down direction (red). The schematics show textures that have particular values for a quantity known as the topological charge: -1 , $+1$, $-\frac{1}{2}$ and $+\frac{1}{2}$ for the skyrmion, antiskyrmion, meron and antimeron, respectively.

called a bimeron, in previous theoretical studies¹².

A more substantial achievement of Yu and colleagues is their demonstration that the square lattice of merons and antimerons could be transformed into a hexagonal lattice of skyrmions by increasing the strength of the applied magnetic field. The authors began with a square lattice of merons and antimerons, whose spins at the core regions pointed in the down and up directions, respectively, in the presence of a perpendicular magnetic field of 20 mT. They increased the field to 60 mT and found that the spin orientations changed completely, and the hexagonal lattice of skyrmions was stabilized.

Finally, by lowering the temperature from 295 K to 120 K, Yu *et al.* compared the stability of the observed spin textures. The authors discovered that the hexagonal lattice of skyrmions is more robust than the square lattice of merons and antimerons, because skyrmions exhibit greater topological protection than do merons and antimerons, in agreement with previous theoretical suggestions¹³. The findings are important because they suggest that many or all kinds of topological spin texture can be realized in a single chiral magnet, and also that a specific spin texture can be selected, depending on the required stability or other characteristics.

Although Yu and colleagues' work is a major step forward in the fields of chiral magnetism and topological spintronics, practical

applications of the observed spin textures could require further breakthroughs. One challenge is that precise control over the crystalline structure and composition of the material is crucial. Such a requirement might limit applications that require high robustness. Moreover, the current-driven behaviour of the textures that is of relevance to spintronics has not yet been observed. Nevertheless, Yu *et al.* have achieved a key experimental discovery that could inspire future engineering efforts in electronic devices that use rich topological spin textures. ■

Seonghoon Woo is at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598, USA.
e-mail: shwoo@ibm.com

1. Yu, X. Z. *et al.* *Nature* **564**, 95–98 (2018).
2. Mühlbauer, S. *et al.* *Science* **323**, 915–919 (2009).
3. Jiang, W. *et al.* *Science* **349**, 283–286 (2015).
4. Fert, A., Reyren, N. & Cros, V. *Nature Rev. Mater.* **2**, 17031 (2017).
5. Nayak, A. K. *et al.* *Nature* **548**, 561–566 (2017).
6. De Alfaro, V., Fubini, S. & Furlan, G. *Phys. Lett. B* **65**, 163–166 (1976).
7. Callan, C. G., Dashen, R. & Gross, D. J. *Phys. Rev. D* **17**, 2717–2763 (1978).
8. Brey, L., Fertig, H. A., Côté, R. & MacDonald, A. H. *Phys. Scripta* **1996**, 154 (1996).
9. Ezawa, M. *Phys. Rev. B* **83**, 100408 (2011).
10. Lin, S.-Z., Saxena, A. & Batista, C. D. *Phys. Rev. B* **91**, 224407 (2015).
11. Ozawa, R. *et al.* *J. Phys. Soc. Jpn* **85**, 103703 (2016).
12. Zhang, X., Ezawa, M. & Zhou, Y. *Sci. Rep.* **5**, 9400 (2015).
13. Nagaosa, N. & Tokura, Y. *Nature Nanotechnol.* **8**, 899–911 (2013).

CELL BIOLOGY

Exit route evolved into entry path in plants

Chloroplast organelles in plant cells are thought to have evolved from bacterial cells. It emerges that the protein–import system in chloroplasts arose from components that export proteins out of bacteria. SEE LETTER P.125

DANNY J. SCHNELL

Chloroplasts are subcellular organelles that provide plants with abundant metabolic capabilities, most notably the ability to capture carbon from atmospheric carbon dioxide through the process of photosynthesis. It has been proposed¹ that the plant kingdom began to evolve when a bacterium, similar to a present-day photosynthetic cyanobacterium, was engulfed by a host cell, and from this bacterial ancestor, chloroplasts eventually arose in the cellular descendants of the host cell. During chloroplast evolution, thousands of genes transferred from the intracellular bacterial genome to the host genome. However, these relocated genes encode proteins that need to be targeted to their site of function within the chloroplast. On page 125, Chen *et al.*² report the identification of a component of the system that imports proteins into chloroplasts. Their finding illuminates how this evolved, and also provides mechanistic insight into how import is coordinated across the two membrane layers that form the chloroplast's outer envelope.

Most chloroplast proteins are made in the cytoplasm. They contain specific amino-acid sequences, termed transit peptides, that are used to direct these proteins from the cytoplasm, across the two membrane layers of the chloroplast and into the interior of the organelle³. Chen and colleagues' work addresses some key questions regarding this protein-import system. The first is, how do the multi-protein complexes, found at the outer and inner membranes of the chloroplast envelope (termed TOC and TIC, respectively) mediate transport in a coordinated way that prevents the mistargeting or misfolding of proteins as they transit through the intermembrane space? Under normal conditions, protein import across the outer and inner membranes seems to occur essentially simultaneously through the TOC and TIC complexes⁴. However, whether there is a physical connection between TOC and TIC, and if so, what its nature is, has been a mystery.

Chen *et al.* report the identification of a previously unknown component of the TIC complex, a protein that they name TIC236, which acts as a link between TIC and TOC. The authors discovered TIC236 using a

biochemical approach to identify proteins that are associated with TOC components. TIC236 is anchored in the inner membrane, where it interacts with components of the TIC complex. Part of TIC236 extends into the intermembrane space, where it interacts with TOC75, a membrane protein that forms part of the channel in the TOC complex (Fig. 1a).

The authors report that, in the plant *Arabidopsis thaliana*, mutations that block the expression of TIC236 are lethal, and mutations that impair TIC236 function reduce the rates of protein import into chloroplasts compared to the import rate in wild-type *A. thaliana*. These results, in addition to the authors' studies of protein–protein interactions, provide compelling evidence that TIC236 provides a key physical link between TIC and TOC. Chen and

colleagues conducted a phylogenetic analysis that provides evidence for the co-evolution of the interacting domains of TOC75 and TIC236, supporting their hypothesis that these proteins evolved as components of interacting complexes. The ability to couple transit through TOC and TIC offers a way of ensuring efficient protein import. This is probably crucial during seedling development, when the rate of protein import into chloroplasts is high, and more than half of the total protein in a cell can be within the chloroplasts⁵.

The second key question this work addresses is, how did the TOC and TIC import systems evolve? Evidence for analogous systems that allow protein import into bacteria is lacking⁶, and, consequently, the evolutionary origin of the chloroplast protein-import system has been an open question. TOC75 is related to the OMP85 family of membrane proteins found on the outer membranes of chloroplasts, energy-generating organelles called mitochondria, and in Gram-negative bacteria, a group that includes the photosynthetic cyanobacteria⁷. Gram-negative bacteria have two membrane layers, and membrane proteins on their surface can be assembled and transported to the cell exterior with the help of protein complexes called TAM or BAM on the outer membrane of the cell. Membrane proteins in these complexes are members of the OMP85 family⁷. It has been proposed⁸ that TOC75 is derived from

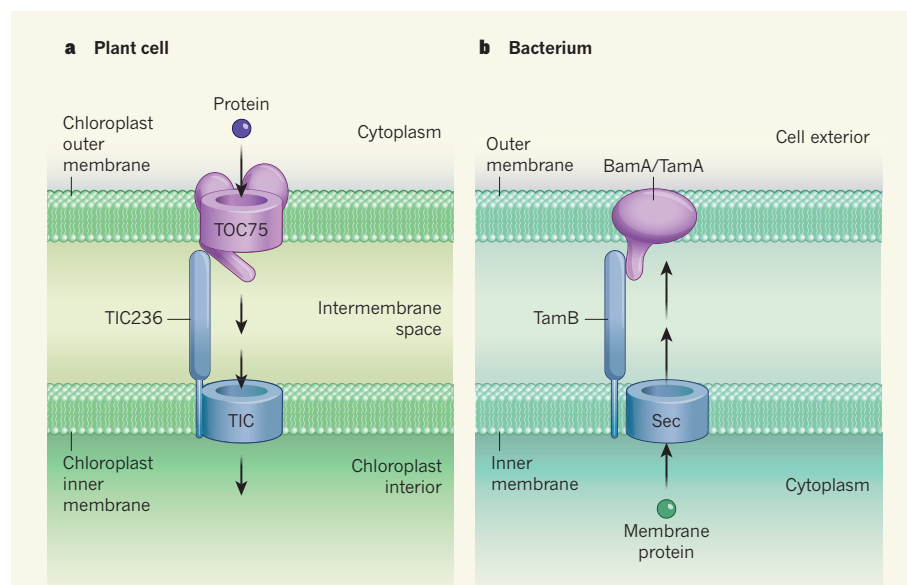


Figure 1 | Protein import into chloroplast organelles evolved from a bacterial protein-export system.

a, Multi-protein complexes called TOC (purple) and TIC (blue), reside respectively on the outer and inner membrane layers of chloroplasts in plant cells. These complexes aid protein import (arrows) into the chloroplast. Chen *et al.*² identified a protein component of TIC, which they termed TIC236, that directly associates with the protein TOC75 in the TOC complex of the plant *Arabidopsis thaliana*. This finding provides insight into how protein transit through both complexes is coordinated. **b**, Bacterial proteins that are part of protein complexes called BAM or TAM aid protein export in the Gram-negative group of bacteria. The authors report that TIC236 is related to the bacterial protein TamB, and it has been proposed⁸ that TOC75 is related to the bacterial proteins BamA or TamA. In Gram-negative bacteria, when membrane proteins are transported from the cytoplasm to the outer membrane, they cross the inner membrane through a protein complex called Sec. TamB is anchored in the inner membrane, and it aids the export of membrane proteins by facilitating⁹ transport between Sec and BamA or TamA proteins on the outer membrane. Chen and colleagues' results suggest that the protein-import system in chloroplasts evolved from this type of protein-export system in the microbial ancestor of chloroplasts.

an ancestral protein related to components of the BAM or TAM systems in the ancestral bacterium that gave rise to the chloroplast.

Chen and colleagues demonstrate that TIC236 is related to TamB, which aids protein transport⁹ between bacterial membrane proteins that form a secretion system (called Sec), located on the inner membrane, and BAM or TAM components in the outer membrane (Fig. 1b). It therefore seems that a mechanism for coupling inner- and outer-membrane transport in chloroplasts has been evolutionarily conserved from an ancestral bacterial system.

However, despite this conservation, the chloroplast protein-import system has evolved to function in the reverse direction relative to the direction of transport in the bacterial export system⁵. The BAM and TAM complexes facilitate export of bacterial proteins from the

cytoplasm to the outer membrane, whereas the TOC and TIC complexes import proteins from outside the chloroplast to inside it. This remarkable reversal of the direction of protein transport probably resulted from the gain of other TOC or TIC proteins that evolved from host-encoded genes to adapt the complexes for the purposes of protein import. These include TOC and TIC receptors and molecular motor proteins known to facilitate transport into the chloroplast³.

Chen and colleagues' results provide convincing evidence for the origin of key elements of the chloroplast protein-import system from an ancestral bacterial protein-export system. Their insights also reveal the adaptation and consequent reversal of an existing protein-targeting pathway that was essential for the ancestral bacteria to successfully take up residence in a host cell, thereby

enabling the host to take advantage of its guest's photosynthetic and metabolic capabilities. ■

Danny J. Schnell is in the Department of Plant Biology, Michigan State University, East Lansing, Michigan 48824, USA.
e-mail: schnelld@msu.edu

1. Archibald, J. M. *Curr. Biol.* **25**, R911–R921 (2015).
2. Chen, Y.-L. et al. *Nature* **564**, 125–129 (2018).
3. Paila, Y. D., Richardson, L. G. L. & Schnell, D. J. *J. Mol. Biol.* **427**, 1038–1060 (2015).
4. Chen, L. & Li, H. *Plant J.* **92**, 178–188 (2017).
5. Ellis, R. J. *Trends Biochem. Sci.* **4**, 241–244 (1979).
6. Day, P. M. & Theg, S. M. *Photosynth. Res.* **138**, 315–326 (2018).
7. Heinz, E. & Lithgow, T. *Front. Microbiol.* **5**, 370 (2014).
8. Day, P. M., Potter, D. & Inoue, K. *Front. Plant Sci.* **5**, 535 (2014).
9. Heinz, E., Selkrig, J., Belousoff, M. J. & Lithgow, T. *Genome Biol. Evol.* **7**, 1628–1643 (2015).

This article was published online on 21 November 2018.

PARTICLE PHYSICS

Long-sought decay of the Higgs boson seen

Measurements of the strength of interactions between the Higgs boson and other particles test the current model of particle physics. A key part of this model has been confirmed by observing the most common decay of the Higgs boson.

BORIS TUCHMING

In 2012, the famous Higgs boson was discovered by the ATLAS and CMS collaborations in proton–proton collisions at the Large Hadron Collider (LHC) at CERN near Geneva, Switzerland^{1,2}. Now, writing in *Physics Letters B*³ and *Physical Review Letters*⁴, the two collaborations report the observation of the Higgs boson decaying to a pair of elementary particles known as bottom quarks. This milestone in particle physics confirms the role of the Higgs field — the quantum field associated with the Higgs boson — in providing particles of matter with mass.

When the standard model of particle physics emerged in the 1960s, the main goal of the ad hoc Higgs field was to explain the masses of the weak vector bosons — the force carriers of the weak nuclear interaction. Mathematical consistency required the force carriers to be massless, whereas the extremely short range of the weak interaction was a signature of massive particles. The Higgs mechanism^{5–8} addressed this issue: the masses of the weak vector bosons are not intrinsic, but are the outcome of interactions between these particles and the all-pervasive Higgs field. It was quickly realized that elementary particles of matter called fermions could also get their masses from interactions with the Higgs field^{9,10}.

Several decades later, twelve elementary fermions are known and are arranged in three families. The first family comprises three charged particles — the up quark, the down quark and the electron — and a neutral particle called the electron neutrino. These fermions are the basic ingredients of ordinary matter: the up and down quarks are the constituents of protons and neutrons, and electron neutrinos are emitted from certain radioactive decays.

For a reason that is not yet fully understood, two replicas of the first family exist. The second family consists of the charm quark, the strange quark, the muon and the muon neutrino, where the charged fermions have

greater masses than their counterparts in the first family. And the third comprises the top quark, the bottom quark, the tau and the tau neutrino, where the charged fermions are even more massive.

After their discovery of the Higgs boson^{1,2}, one objective of the ATLAS and CMS collaborations was to probe the particle's properties, such as its couplings to fermions — the strength of its interactions with fermions. In the current papers, the collaborations combined all the data that they recorded between 2011 and 2017, and each claims to have observed the decay of the Higgs boson to bottom quarks.

In both sets of data, the decay signal is larger than the background, which arises from other particle-physics processes. The statistical significance of the signal is 5.4 and 5.6 standard deviations for the ATLAS and CMS experiments, respectively — well above the conventional threshold of 5 standard deviations needed to claim observation. In addition, the overall yields of the decay are in agreement with standard-model predictions within an experimental uncertainty of roughly 20%.

The Higgs boson decays almost immediately after it is produced. The probability that a particular decay will occur depends on the

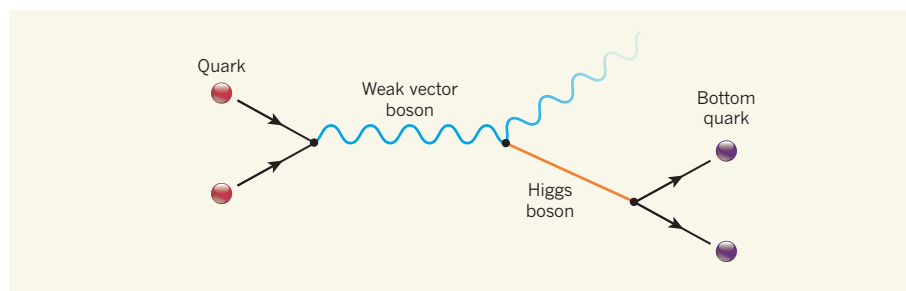


Figure 1 | Production of the Higgs boson together with a weak vector boson. The ATLAS³ and CMS⁴ collaborations report evidence that a particle known as the Higgs boson can decay to pairs of elementary particles called bottom quarks. To detect this decay, the collaborations looked for a particular process in which two quarks arising from colliding protons fuse to form a weak vector boson — a force carrier of the weak nuclear interaction. The weak vector boson emits a Higgs boson that decays to bottom quarks.

couplings to the Higgs boson, which are determined by the masses of the decay products. Because bottom quarks are among the heaviest fermions, the decay to these particles is the most common, occurring about 58% of the time. But even though this decay is dominant, in proton–proton collisions the signal is overwhelmed by the background of bottom quarks produced by the strong nuclear interaction. For this reason, the discovery of the Higgs boson in 2012 involved decays only to vector bosons: photons from the electromagnetic interaction and weak vector bosons from the weak interaction.

To observe the decay to bottom quarks, the two collaborations had to look for subdominant modes of Higgs–boson production, such as the production of the Higgs boson together with a weak vector boson (Fig. 1). A deep understanding of the responses of the particle detector, and sophisticated data-analysis methods that included machine learning, were needed to precisely reconstruct the energies and momenta of the weak vector bosons, tag the jets of particles arising from the bottom quarks, model all the backgrounds and separate these backgrounds from the signal.

The findings are not entirely surprising, for at least two reasons. First, there have been several pieces of evidence for the decay of the Higgs boson to bottom quarks in the past. In 2012, a signal at the level of 2.8 standard deviations was claimed by scientists at the Tevatron proton–antiproton collider, located near Chicago¹¹. Between 2012 and 2018, the ATLAS and CMS collaborations regularly reported outcomes of their search for the decay. In their latest papers before the current work, they obtained evidence at the level of 3.6 and 3.8 standard deviations, respectively^{12,13}. These different pieces of evidence could be considered as a combined observation of the decay.

Second, many other experimental results at the LHC are constraining what could actually be observed regarding this decay. For example, if the Higgs boson had behaved as in the standard model, but had had zero coupling to bottom quarks, the yields of all the other decay modes would be enhanced by a factor of about 2.4, which is contradicted by the data. Considering the overall picture, unless there exist unexpected cancelling effects, the allowed deviations from the standard model are at the level of a few per cent — below the current 20% sensitivity of experiments at the LHC.

Nevertheless, the current results are a great achievement and constitute a major milestone in particle physics. Together with observations earlier this year of the Higgs boson decaying to tau particles¹⁴ and the production of the Higgs boson together with top quarks^{15,16}, the findings directly establish interactions between the Higgs boson and the third family of fermions, therefore pointing to the Higgs field as the origin of fermion masses.

The results are the starting point of an era of precision measurement for the couplings of the Higgs boson to fermions. With more data from the LHC — in particular, after upgrades to the beam intensity in a few years — an accuracy of a few per cent in the measurements should be obtained. This would open the possibility of finding deviations from the standard model and of, for example, uncovering currently unknown particles.

Another milestone would be observing the couplings of the Higgs boson to the second family of fermions. The decay of the Higgs boson to a pair of muons is within the reach of the future upgraded LHC. However, because of the extremely high background in proton–proton collisions, the decay to charm quarks could probably be demonstrated only by using a giant electron–positron collider, which is yet to be constructed. The Higgs boson is therefore far from having revealed all of its secrets. ■

Boris Tuchming is at the Institute of Research into the Fundamental Laws of the Universe and in the Department of Particle

Physics, CEA University of Paris-Saclay, 91191 Gif-sur-Yvette, France.
e-mail: boris.tuchming@cea.fr

1. ATLAS Collaboration. *Phys. Lett. B* **716**, 1–29 (2012).
2. CMS Collaboration. *Phys. Lett. B* **716**, 30–61 (2012).
3. ATLAS Collaboration. *Phys. Lett. B* **786**, 59–86 (2018).
4. CMS Collaboration. *Phys. Rev. Lett.* **121**, 121801 (2018).
5. Higgs, P. W. *Phys. Lett.* **12**, 132–133 (1964).
6. Englert, F. & Brout, R. *Phys. Rev. Lett.* **13**, 321–323 (1964).
7. Guralnik, G. S., Hagen, C. R. & Kibble, T. W. B. *Phys. Rev. Lett.* **13**, 585–587 (1964).
8. Kibble, T. W. B. *Phys. Rev.* **155**, 1554–1561 (1967).
9. Weinberg, S. *Phys. Rev. Lett.* **19**, 1264–1266 (1967).
10. Salam, A. *Elementary Particle Theory* (ed. Svartholm, N.) 367–387 (Almqvist & Wiksell, 1968).
11. CDF and D0 Collaborations. *Phys. Rev. Lett.* **109**, 071804 (2012).
12. ATLAS Collaboration. *J. High Energy Phys.* **2017** (12), 24 (2017).
13. CMS Collaboration. *Phys. Lett. B* **780**, 501–532 (2018).
14. CMS Collaboration. *Phys. Lett. B* **779**, 283–316 (2018).
15. ATLAS Collaboration. *Phys. Lett. B* **784**, 173–191 (2018).
16. CMS Collaboration. *Phys. Rev. Lett.* **120**, 231801 (2018).

This article was published online on 21 November 2018.

METABOLISM

Reducing oxygen consumption in fat

Low oxygen levels are a hallmark of expanding fat tissue in obesity, and can lead to type 2 diabetes. In addition to a lack of adequate blood supply, increased oxygen demand in fat cells now emerges as being key to this harmful state.

NOLWENN JOFFIN & PHILIPP E. SCHERER

A major cause of type 2 diabetes is obesity, in which fat cells expand rapidly, in both size and number, and their oxygen demand outstrips supply. This low-oxygen state, known as hypoxia, leads to upregulation of the anti-hypoxic protein HIF-1 α , which in turn causes tissue inflammation and prevents fat cells (adipocytes) from responding normally to insulin^{1,2}. Hypoxia in expanding fat is often thought of mainly as a problem of supply, caused by the inability of blood vessels that deliver oxygen to grow as fast as the surrounding tissue^{3,4}. Writing in *Nature Metabolism*, Seo *et al.*⁵ highlight a pathway by which excessive oxygen consumption in adipocytes can also contribute to hypoxia in expanding fat tissue. This pathway involves the enhanced activity of the enzyme adenine nucleotide translocase 2 (ANT2) in energy-generating organelles called mitochondria.

During normal mitochondrial respiration, electrons are transferred between a series of molecules, and this transfer is coupled to the

removal of hydrogen ions (H⁺, also known as protons) from the central matrix of the mitochondrion into the space between its outer and inner membranes. This process creates a proton gradient that drives the production of energy-carrying ATP molecules in mitochondria by the enzyme ATP synthase. But the process can become uncoupled if protons leak across the inner mitochondrial membrane. Uncoupled respiration results in inefficient ATP production, and thereby increases the intracellular demand for oxygen for further respiration.

High levels of uncoupled respiration can alter cellular physiology, and inhibiting uncoupled respiration with various compounds increases cellular oxygen levels, decreasing hypoxia and so reducing HIF-1 α levels⁶. Any manipulation that leads to a decrease in cellular HIF-1 α activity in fat is metabolically beneficial¹. Thus, a better understanding of uncoupled respiration and how to manipulate it is desirable.

Previous work⁷ by the group that carried out the current study has shown that the rate of

oxygen consumption in the white adipocytes of mice increases if the animals eat a high-fat diet. The group proposed that increased levels of circulating free fatty acids in the blood of obese animals led to activation of ANT2. Excessive ANT2 activity results in an increased proton leak back into the mitochondrion⁸, leading to elevated levels of uncoupled mitochondrial respiration.

Seo *et al.* have now developed a mouse model in which expression of the *Ant2* gene is lowered specifically in adipocytes, enabling them to provide proof of this mechanism in the current study. First, the authors showed that the mutant mice became as obese as wild-type mice when fed a high-fat diet, with no difference in total body weight or physical activity between the two groups. However, an increase in adipocyte size (hypertrophy) led to higher fat-tissue weight in *Ant2*-mutant mice than in controls. Despite the well-documented association between adipocyte hypertrophy and hypoxia, the authors found that intracellular oxygen tension — a measure of the concentration of oxygen in the cell, which is decreased in hypoxia — was higher in *Ant2*-mutant mice than in controls. The group showed that the maintenance of oxygen tension was attributable to a decrease in oxygen consumption, rather than to changes in oxygen supply or blood-vessel density, suggesting that ANT2 is a crucial determinant of the rate at which adipocytes consume oxygen in obese animals.

The improved adipocyte oxygen tension in the *Ant2*-mutant mice was independent of ATP synthase, indicating that it did not relate to changes in coupled mitochondrial respiration. Instead, Seo and colleagues confirmed that the mutation led to a decrease in the leakage of protons across the inner mitochondrial membrane that increased the electrical potential across the membrane. This, in turn, enabled more-efficient energy production and less-uncoupled respiration (Fig. 1), and so improved adipocyte survival.

A range of immune cells are recruited to expanded fat tissue, triggering inflammatory responses and tissue scarring known as fibrosis. But Seo *et al.* showed that the functional improvement in mitochondria caused by adipocyte-specific *Ant2* depletion reduces this response — an improvement that is also seen if vascular density is increased in fat tissue through genetic engineering⁴. As expected, this decrease in inflammation and fibrosis led to improved glucose tolerance and enhanced insulin sensitivity in the livers of the *Ant2*-mutant mice. Moreover, the researchers showed that depletion of *Ant2* in the adipocytes of mice that have already developed glucose intolerance and insulin resistance can reverse these effects.

These findings are of interest for several reasons. First, many studies have emphasized the need for adequate vascularization in fat to prevent hypoxia⁹. But Seo and co-workers put the adipocyte centre stage as a driving force for hypoxia, highlighting how a defect

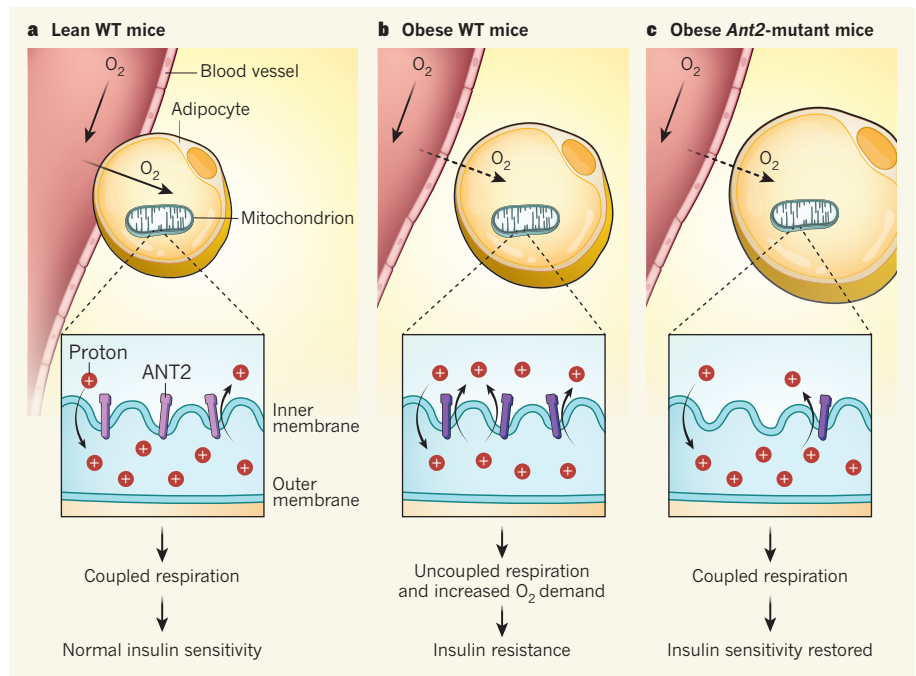


Figure 1 | The ANT2 enzyme in obesity. **a**, Fat cells (adipocytes) receive oxygen from surrounding blood vessels for coupled respiration, in which protons (positively charged hydrogen ions) are removed from the centre of organelles called mitochondria into the space between the inner and outer membranes, generating a membrane potential that drives energy production. The enzyme ANT2 causes proton leakage back into the organelle. This can lead to less-efficient, uncoupled respiration, but ANT2 activity is low in lean wild-type (WT) mice. Coupled respiration ensures normal insulin sensitivity in fat in these animals. **b**, In obese WT mice, adipocytes become larger and receive less oxygen (dashed arrow) owing to lack of an adequate blood supply. Seo *et al.*⁵ report that, in addition, oxygen demand increases because ANT2 activity is increased in obese animals (indicated by darker colour), which lowers membrane potential and drives uncoupled respiration. This leads to insulin resistance (a hallmark of diabetes) in the surrounding tissue. **c**, Reducing adipocyte expression of the *Ant2* gene in obese mice decreases ANT2 levels, lowers the rate of uncoupled respiration and therefore decreases oxygen demand, and so restores insulin sensitivity. (Surprisingly, the adipocytes of these mutant animals are larger than those of obese WT mice, but display higher insulin sensitivity.)

in the fat cell that leads to intracellular oxygen depletion can drive much broader metabolic changes. Second, the authors' *Ant2*-deficient mice show an overall increase in adipocyte size. This finding is counter-intuitive, because adipocyte hypertrophy is generally associated with defective metabolism — this observation therefore needs further investigation. One possible explanation is that reduced HIF-1 α levels in the hypertrophic cells promote their survival. Third, the researchers demonstrate that intracellular oxygen tension is higher in the fat of people who have metabolically normal obesity than in those with metabolically abnormal obesity. This is in line with the respective insulin sensitivities of these conditions, indicating that the authors' findings might have clinical relevance.

Seo and colleagues' work defines modulation of ANT2 as a potential strategy to improve systemic metabolic defects, including type 2 diabetes. Combined with the fact that ANT2 has been suggested to be an attractive anti-cancer target¹⁰, this makes ANT2 modulators prime candidates for drug development. This is even more appealing in light of the fact that Seo *et al.* only partly inhibited *Ant2* expression in the current study, rather than completely

deleting the gene. Thus, small-molecule drugs, which can only partially block the activity of their target enzyme, might provide the desired effects. Efforts to identify such inhibitors should prove rewarding in the future. ■

Nolwenn Joffin and Philipp E. Scherer are in the Touchstone Diabetes Center, Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA.
e-mail: philipp.scherer@utsouthwestern.edu

1. Sun, K., Halberg, N., Khan, M., Magalang, U. J. & Scherer, P. E. *Mol. Cell. Biol.* **33**, 904–917 (2013).
2. Halberg, N. *et al.* *Mol. Cell. Biol.* **29**, 4467–4483 (2009).
3. An, Y. A. *et al.* *eLife* **6**, e24071 (2017).
4. Sun, K. *et al.* *Proc. Natl Acad. Sci. USA* **109**, 5874–5879 (2012).
5. Seo, J. B. *et al.* *Nature Metab.* <https://doi.org/10.1038/s42255-018-0003-x> (2018).
6. Hagen, T., Taylor, C. T., Lam, F. & Moncada, S. *Science* **302**, 1975–1978 (2003).
7. Lee, Y. S. *et al.* *Cell* **157**, 1339–1352 (2014).
8. Chevrollier, A., Loiseau, D., Reynier, P. & Stepien, G. *Biochim. Biophys. Acta* **1807**, 562–567 (2011).
9. Crewe, C., An, Y. A. & Scherer, P. E. *J. Clin. Invest.* **127**, 74–82 (2017).
10. Sharaf el dein, O., Mayola, E., Chopineau, J. & Brenner, C. *Curr. Drug Targets* **12**, 894–901 (2011).

This article was published online on 19 November 2018.

ENVIRONMENTAL SCIENCE

Ammonia maps make history

Ammonia emissions harm humans and the environment. An analysis shows that satellites can locate sources precisely, and could thus help to monitor compliance with international agreements to limit such emissions. [SEE LETTER P.99](#)

MARK A. SUTTON & CLARE M. HOWARD

According to the tenth-century Arabic geographer Al-Masudi, travellers through the mountains between Samarkand and China had to pass through a valley where the smoke was so dense that the Sun's rays could not penetrate. Al-Masudi recorded¹ how paid porters would “use sticks to drive the passengers on their journey; for any stoppage or rest would be fatal to the traveller, in consequence of the irritation which the ammoniacal vapours of this valley produce on the brain, and on account of the heat”. His graphic account describes the earliest known industrial source of ammonia emissions, and has fresh significance in light of a study reported on page 99 by Van Damme and colleagues². The authors have mapped atmospheric ammonia levels with unprecedented precision around the globe, and have quantified emissions from this ancient source for the first time — along with those from a host of previously uncharacterized industrial and agricultural hotspots.

In the valley mentioned by Al-Masudi, locals were exploiting the spontaneous natural combustion of surface coal seams. They used stone huts to collect ammonium chloride and other ammonium salts^{3,4} carried by the fumes, with the remaining emissions contributing to air pollution. Although this oldest of ammonium industries is no longer in business, Van Damme *et al.* identify two sites, at Abakan in Russia and Jharia in India (Fig. 1), that are emitting ammonia from burning coal mines. Their demonstration that global satellite observations can now detect such ammonia sources represents a historic moment for science.

Al-Masudi's example makes it clear why we should care about ammonia. Emitted ammonia reacts rapidly with other air pollutants, and thereby helps to form fine particulate matter that shortens the human lifespan through respiratory and coronary diseases⁵. Moreover, gaseous ammonia and ammonium compounds formed from it in the atmosphere are deposited into ecosystems, damaging sensitive habitats — especially those naturally adapted to need clean air. Ammonia emissions from agricultural sources also reduce the efficiency with which nitrogen is used though the food-production chain, which has knock-on

consequences that increase greenhouse-gas emissions and contribute to water pollution⁶.

Van Damme *et al.* carried out a high-spatial-resolution analysis based on nine years of data derived from the Infrared Atmospheric Sounding Interferometer (IASI) — an instrument that takes twice-daily measurements of atmospheric ammonia levels — on the Metop-A meteorological satellite. This allowed the researchers to estimate ammonia emissions from 248 hotspots (defined by the authors as areas with diameters of less than 50 kilometres) and a further 178 regional sources (which have no clearly defined hotspot).

This is not the first report of ammonia distribution mapped from satellite observations. Earlier publications used IASI data⁷ or measurements from other infrared-observing platforms^{8,9} to produce global maps and characterize source regions. What sets Van Damme and colleagues' analysis apart is the comprehensiveness and diversity of quantified ammonia sources. The study shows how satellite technology is coming of age as it starts to fulfil multiple scientific and policy-assessment objectives.

For example, the authors used a method called oversampling to produce a much more precise global map than was previously available. The IASI instrument scans the entire globe

daily, and records observations at each observation point at 09:30 and 21:30 hours (local time), but always at slightly different positions. By averaging the nine-year record, the researchers were able to derive fine-scale maps with a resolution of $0.01^\circ \times 0.01^\circ$, from which they identified hundreds of ammonia sources. They also used a simple ‘inversion’ method to estimate the emissions produced from all of the sources. The authors complemented these approaches by using aerial photographs and other independently obtained data to help them characterize the nature of the hotspots.

Most of the hotspots were found to be associated with intensive livestock farming and industrial activities. Van Damme *et al.* also discovered a previously unknown natural source — Lake Natron in Tanzania, where drying mudflats are found to release substantial amounts of ammonia to the atmosphere.

One of the authors' key findings is that many of the industrial sources are missing from ‘bottom-up’ inventories of ammonia emissions (which combine data on the intensity of ammonia-emitting activities with estimated quantities called emission factors). This shows that satellite technology can now be used as an auditing tool for the national reporting of ammonia emissions. Some countries might not report emissions from specific sources because, for example, national regulations do not require polluting activities to be registered. The use of standard emission factors in bottom-up inventories might also lead to errors in national reporting, emphasizing the need for independent verification methods.

As with any emerging technology, some limitations of satellite monitoring remain to be overcome. The largest of these are the requirement for the atmosphere to be cloud-free when measurements are made, and the need for a sizeable temperature difference between land or sea surface and the atmosphere — which limits the zones at which measurements can usefully



Figure 1 | Burning coalfields at Jharia, India.

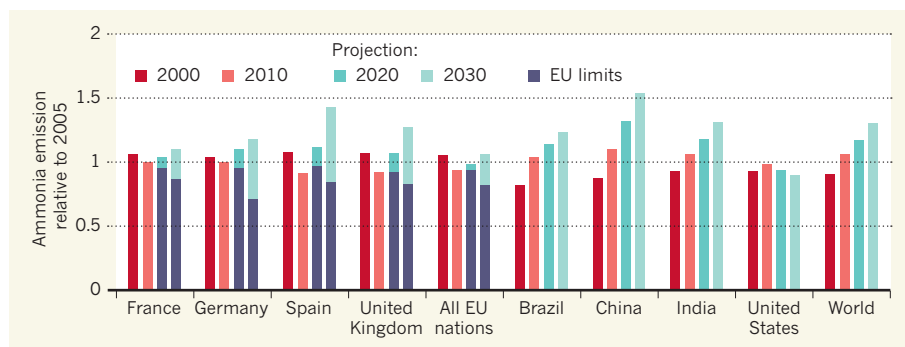


Figure 2 | Past and projected ammonia emissions. Ammonia emissions from many countries are on course to increase in the next couple of decades — including in the European Union, where several countries are set to exceed legally binding limits for 2020 and 2030. Data are shown relative to ammonia emissions in 2005 for each country. Graphs for the European countries are plotted from official EU data (see go.nature.com/2awg8sc), whereas non-EU data are from the EDGAR 4.3 inventory¹¹. Projected future values are extrapolated from the most recent five years of available data. Both data sources are ‘bottom-up’ inventories, which combine data on the intensity of ammonia-emitting activities with estimated quantities called emission factors. Van Damme *et al.*² show that satellites now have the capability to help assess compliance with legally binding limits.

be made to warm-temperate and tropical climates. There is also huge potential to improve on Van Damme and colleagues’ inversion technique, which assumes that the atmospheric lifetime of ammonia is constant everywhere. This simplification will underestimate ammonia emissions at sources in windy locations, such as on coasts or in mountain areas. Together, these limitations might explain why the authors do not detect high ammonia levels at any seabird colonies, which are known to be substantial ammonia hotspots, especially in sub-polar regions¹⁰. Curiously enough, the authors identify several fire-based sources (including the

second-largest global ammonia hotspot, found in West Africa), but exclude many of these from their detailed analysis.

Perhaps the most important feature of the new analysis, however, is that it has demonstrated ammonia trends at specific locations. For example, the authors detected changes of 15–20% in ammonia levels at hotspots over the period of the study (see Fig. 4 of the paper²). The achievement of this accuracy for hotspots implies that even better precision could probably be achieved for observations at national and regional scales.

Achieving this capability now is especially

timely. Ammonia emissions in many countries are currently increasing (Fig. 2), even in the European Union, which has committed to achieving an overall reduction of 6% by 2020 and 19% by 2030, compared with 2005 levels (see Annex II at go.nature.com/2e2gphe). Combined with atmospheric models (which are necessary for considering the effects of ammonia’s interactions with acidic gases and particulate matter), satellite technology offers a valuable independent tool with which to check whether countries are really achieving their goals. ■

Mark A. Sutton and Clare M. Howard are at the International Nitrogen Management System, NERC Centre for Ecology and Hydrology, Edinburgh Research Station, Edinburgh EH26 0QB, UK.
e-mails: ms@ceh.ac.uk; cbritt@ceh.ac.uk

1. El-Mas’ūdū’s *Historical Encyclopaedia Entitled “Meadows of Gold and Mines of Gems”* Vol. I (transl. Sprenger, A.) 360 (Oriental Translation Fund, 1841).
2. Van Damme, M. *et al. Nature* **564**, 99–103 (2018).
3. Belakovski, D. *Lapis* **15**, 21–26 (1990).
4. Sutton, M. A., Erismann, J. W., Dentener, F. & Möller, D. *Environ. Pollut.* **156**, 583–604 (2008).
5. Brunekreef, B. *et al. Lancet Respir. Med.* **3**, 831–832 (2015).
6. Sutton, M. A. *et al. Nature* **472**, 159–161 (2011).
7. Clarisse, L., Clerbaux, C., Dentener, F., Hurtmans, D. & Coheur, P.-F. *Nature Geosci.* **2**, 479–483 (2009).
8. Shephard, M. W. & Cady-Pereira, K. E. *Atmos. Meas. Tech.* **8**, 1323–1336 (2015).
9. Warner, J. X. *et al. Geophys. Res. Lett.* **44**, 2875–2884 (2017).
10. Riddick, S. N. *et al. Atmos. Environ.* **184**, 212–223 (2018).
11. Crippa, M. *et al. Earth Syst. Sci. Data* **10**, 1987–2013 (2018).

the pancreas begins to form, these progenitor cells are segregated into domains that give rise to specific cell lineages⁵.

One domain forms the cells that make digestive enzymes, and the other domain, termed the trunk domain, develops from cells called bipotent pancreatic progenitors (bi-PPs) that can give rise to two cell types (Fig. 1), pancreatic duct cells and hormone-producing cells^{4,5}. A hallmark of the bi-PP cells that will differentiate into hormone-producing cells^{3,4}, such as β -cells, is the expression of the transcription factor NGN3.

Mamidi and colleagues investigated what determines the type of cell that develops from a bi-PP cell. They used experimental systems that included *in vivo* mouse models and *in vitro* studies of organ cultures and human embryonic stem cells that had differentiated to form pancreatic cells.

In some of the *in vitro* studies, the authors used micropatterned glass slides to restrict the location and shape of the regions to which stem cells could attach and, hence, grow on. This revealed that cells confined to a limited space were more likely to differentiate into hormone-producing cells, and that cells that could spread out over a large area were less likely to form this type of cell.

DEVELOPMENTAL BIOLOGY

Location matters for insulin-producing cells

Intrinsic and extrinsic cues drive dynamic processes that control cell fate during organ development. A study of mouse and human cells reveals how these inputs affect cells that make the essential hormone insulin. SEE LETTER P.114

FRANCESCA M. SPAGNOLI

During development, cells proliferate and differentiate to enable organs to achieve their final functional architecture¹. As cells develop to reach their mature state, they respond to various extrinsic cues provided by the surrounding microenvironment and acquire a fate that can be determined by their location in a tissue. But little is known about how these cues drive intracellular changes, such as transcription or differentiation, or how tissue architecture and cellular rearrangements can, in turn, affect cell fate. On page 114, Mamidi *et al.*² provide insight into how cell location

and exposure to certain external cues can affect whether cells in the developing pancreas give rise to β -cells that make the protein insulin. Deficiencies in insulin-producing cells can lead to diabetes, so a better understanding of how these cells form could have clinical implications.

As well as the hormone insulin, which has an essential role in the regulation of blood-glucose levels³, the pancreas produces digestive enzymes. The organ develops by a complex stepwise process that gives rise to many cell types^{3,4}. Embryonic pancreatic cells, also termed progenitor cells, initially express the transcription factor PDX1 and can generate all of the cell types found in the pancreas^{3,5}. When

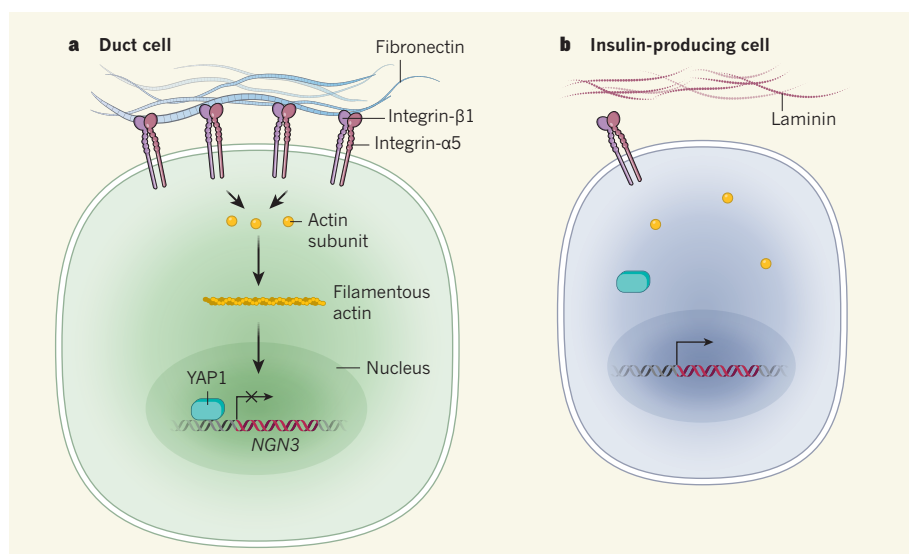


Figure 1 | Extracellular cues affect the formation of insulin-producing cells. Mamidi *et al.*² used mouse and human cells to investigate what determines whether a type of cell in the developing pancreas — called a bipotent pancreatic progenitor — becomes a duct cell or a cell that makes the protein insulin. **a**, The authors report that duct-cell formation is associated with the presence of the protein fibronectin in the cell's microenvironment. Cellular interaction with fibronectin can be mediated by the proteins integrin- α 5 and integrin- β 1, which form a fibronectin receptor. This interaction is associated with cellular spreading (an increase in surface area). In such cells, the protein actin is in a filamentous form and there are high levels of the protein YAP1 in the nucleus. YAP1 represses the transcription of the gene encoding the protein NGN3, which is characteristic of the progenitor cells that will give rise to insulin-producing cells. **b**, The authors found that the formation of insulin-producing cells is associated with the presence of the protein laminin in the cellular microenvironment. Compared with duct cells, insulin-producing cells have lower levels of integrin- α 5 (and therefore lower levels of fibronectin receptors), higher expression of the gene encoding NGN3, and less filamentous actin, cellular spreading and nuclear YAP1.

How might differences in cellular shape affect cell fate? The authors tested whether the protein YAP1 might be involved. YAP1 is a component of the Hippo signalling cascade, which controls organ size, and can function as a sensor of cell shape and a mediator of cellular responses to mechanical stimuli⁶. In the cells cultured on a micropatterned surface, cellular spreading was associated with sustained nuclear activity of YAP1 and low levels of PDX1 and NGN3, whereas cell confinement corresponded to low levels of YAP1 in the nucleus and high levels of PDX1 and NGN3. The authors then engineered mice that lack YAP1 in their pancreatic progenitor cells. The number of insulin-expressing cells in these animals was higher than that in mice that had normal YAP1 expression, providing *in vivo* results that are consistent with their *in vitro* work.

Mamidi *et al.* sought to define the downstream targets of YAP1 in pancreatic progenitor cells. They found that YAP1 regulates the signalling pathway that is mediated by the protein Notch, and also that YAP1 directly represses transcription of the gene that encodes NGN3.

The authors also wanted to understand the pathway upstream of YAP1. How does an extracellular cue regulate YAP1, and does the same cue trigger the generation of hormone-producing cells? Changes in the arrangement of the protein actin in cells can affect YAP1 activity and cause changes in cell shape⁶. Using human embryonic stem cells that had differentiated

into pancreatic cells and mouse pancreatic tissue grown *in vitro*, Mamidi *et al.* showed that blocking the assembly of actin into filaments resulted in cells having a reduced surface area, low nuclear YAP1 levels and high NGN3 levels, compared with cells in which filamentous actin assembly had not been perturbed. It remains to be determined exactly how actin regulation converges on a YAP1-mediated signalling pathway to govern cell fate.

The most interesting aspect of Mamidi and colleagues' work is undoubtedly their finding that YAP1 activity in a cell responds to external cues from the pancreatic microenvironment that affect the assembly of actin filaments. To investigate how such external cues might affect actin and YAP1, the authors focused on integrin and FAK proteins, which function at a 'crossroads' between filamentous actin and the surrounding microenvironment. These proteins can mediate the interactions between cells and the extracellular matrix material in their surroundings⁶. The authors report that a high level of integrin- α 5 in mouse or human cells correlates with cells that are in a bi-PP state or cells that are duct progenitors, whereas low levels of this integrin are associated with the formation of hormone-producing cells. Integrin- α 5 and integrin- β 1 can form a receptor for the extracellular protein fibronectin⁷; therefore, a change in integrin expression might affect the ability of a pancreatic progenitor cell to respond to the extracellular environment in a way that affects cell fate.

The authors' studies of mouse and human cells show that the presence of extracellular fibronectin promotes cellular spreading, whereas exposure to an extracellular matrix protein called laminin inhibits spreading and is accompanied by a reduction in the level of integrin- α 5 and in the level of YAP1 in the nucleus. Perhaps a feedback loop based on signalling between extracellular-matrix material and integrin receptors exists in which pancreatic progenitor cells that are exposed to laminin have low expression of integrin- α 5, lose the ability to respond to fibronectin, and develop into insulin-producing cells. But how the extracellular matrix might control integrin expression is an open question.

Mamidi *et al.* suggest that regions of different extracellular-matrix composition in the developing pancreas microenvironment might exert different effects on progenitor cells, thereby influencing cell fate. The authors characterized the distribution of fibronectin and laminin in the developing mouse pancreas, and noted that the cells that will develop into insulin-producing cells are more commonly found in association with laminin than with fibronectin. However, these developing insulin-producing cells are also exposed to some fibronectin, suggesting that cells might have to respond to complex gradients of extracellular-matrix material *in vivo*.

The authors' analysis of interactions between the extracellular matrix and cells through static 'snapshots' of the process offers a valuable starting point. However, it would be even better to understand the dynamics of the interactions in time and space in an *in vivo* context, given that the deposition of extracellular-matrix material and cellular positions change constantly during development. It would also be fascinating to discover which cells deposit the extracellular material that surrounds the developing pancreas.

Mamidi and colleagues' work might have direct implications for efforts to generate β -cells for cell-replacement therapies to treat diabetes. Their study suggests avenues of investigation for improving the strategies used to coax human embryonic stem cells to differentiate into insulin-producing cells⁸. ■

Francesca M. Spagnoli is at the Centre for Stem Cells and Regenerative Medicine, Faculty of Life Sciences & Medicine, King's College London, London SE1 9RT, UK.
e-mail: francesca.spagnoli@kcl.ac.uk

- Hogan, B. L. M. *Cell* **96**, 225–233 (1999).
- Mamidi, A. *et al.* *Nature* **564**, 114–118 (2018).
- Shih, H. P., Wang, A. & Sander, M. *Annu. Rev. Cell Dev. Biol.* **29**, 81–105 (2013).
- Pan, F. C. & Wright, C. *Dev. Dyn.* **240**, 530–565 (2011).
- Zhou, Q. *et al.* *Dev. Cell* **13**, 103–114 (2007).
- Totaro, A., Panciera, T. & Piccolo, S. *Nature Cell Biol.* **20**, 888–899 (2018).
- Huveneers, S. & Danen, E. H. J. *J. Cell Sci.* **122**, 1059–1069 (2009).
- Sneddon, J. B. *et al.* *Cell Stem Cell* **22**, 810–823 (2018).

This article was published online on 28 November 2018.

Change in future climate due to Antarctic meltwater

Ben Bronselaer^{1,2,3*}, Michael Winton², Stephen M. Griffies^{2,3}, William J. Hurlin², Keith B. Rodgers³, Olga V. Sergienko^{2,3}, Ronald J. Stouffer^{1,2} & Joellen L. Russell¹

Meltwater from the Antarctic Ice Sheet is projected to cause up to one metre of sea-level rise by 2100 under the highest greenhouse gas concentration trajectory (RCP8.5) considered by the Intergovernmental Panel on Climate Change (IPCC). However, the effects of meltwater from the ice sheets and ice shelves of Antarctica are not included in the widely used CMIP5 climate models, which introduces bias into IPCC climate projections. Here we assess a large ensemble simulation of the CMIP5 model ‘GFDL ESM2M’ that accounts for RCP8.5–projected Antarctic Ice Sheet meltwater. We find that, relative to the standard RCP8.5 scenario, accounting for meltwater delays the exceedance of the maximum global-mean atmospheric warming targets of 1.5 and 2 degrees Celsius by more than a decade, enhances drying of the Southern Hemisphere and reduces drying of the Northern Hemisphere, increases the formation of Antarctic sea ice (consistent with recent observations of increasing Antarctic sea-ice area) and warms the subsurface ocean around the Antarctic coast. Moreover, the meltwater-induced subsurface ocean warming could lead to further ice-sheet and ice-shelf melting through a positive feedback mechanism, highlighting the importance of including meltwater effects in simulations of future climate.

Observations have shown an acceleration in mass loss from the Antarctic Ice Sheet in recent years^{1–3}. A recent study predicts that the Antarctic Ice Sheet will contribute almost 1 m to global sea-level rise by the end of the twenty-first century under the RCP8.5 scenario⁴. However, the latest generation of climate models included in the Coupled Model Intercomparison Project Phase 5 (CMIP5)⁵ for the fifth IPCC assessment report do not account for ice melt in future climate projections. Although output from model simulations is used to project sea-level rise due to ice-sheet mass loss, feedback on the climate system is not included. In addition, ice-sheet and ice-shelf melt will not be accounted for in the upcoming CMIP6 standard suite of experiments⁶.

Although the effect of the Antarctic Ice Sheet is most often considered in terms of global sea level^{4,7}, idealized climate model simulations show distinct effects of the meltwater flux on the simulated climate. In the Southern Ocean, the water mass stratification is such that a cold surface layer lies above a deep warm layer (called circumpolar deep water, CDW). Ice-sheet meltwater reduces ocean mixing and further isolates the warm CDW from the surface, which has been shown to reduce sea-surface temperatures and cause subsurface ocean warming around Antarctica^{8–10}, and to potentially increase sea-ice extent^{11,12}. This mechanism suggests that the cooling influence of meltwater released into the Southern Ocean can offset some of the twenty-first-century warming, delaying the exceedance of the 2015 United Nations Climate Change Conference (COP21) maximum warming targets¹³. Elevated subsurface ocean temperatures offshore can propagate into ice-shelf cavities and increase basal melting of ice shelves^{11,14–17}, although modelling studies disagree on the magnitude and effect of the climate response to Southern Ocean freshening^{18–20}. In addition, this mechanism could alter atmospheric heat transport and rainfall patterns^{21,22}.

Here, we assess the effect of Antarctic meltwater on the climate state in a CMIP5 climate model simulation that accounts for the historical

and RCP8.5-projected ice-sheet meltwater⁴ (Extended Data Fig. 1). We show that modifications to a climate projection can be substantial, indicating a bias in current IPCC climate models. In particular, we use a large ensemble simulation of the CMIP5 model GFDL ESM2M (Geophysical Fluid Dynamics Laboratory’s Earth system model version 2M) between 1950 and 2100 (Methods). We refer to the simulations that include the effect of ice-sheet meltwater as the ‘meltwater ensemble’ and the simulations without meltwater as the ‘standard ensemble’. By using a large ensemble simulation, we robustly quantify the statistical significance of the effects of meltwater on important aspects of the simulated climate and the time when the meltwater ensemble diverges from the standard ensemble.

To characterize the climate response to the release of Antarctic meltwater, we assess four key fields that are routinely used in IPCC assessments of climate change: surface air temperature (SAT), precipitation, sea-ice cover and the ocean temperature around Antarctica at a depth of 400 m. We include the subsurface ocean temperature to highlight a possible feedback on ice-shelf melting. A depth of 400 m is representative for assessing the effect of ocean temperature on ice shelves because it is the typically observed depth of warm CDW from which intrusions onto the continental shelf are sourced⁴. For each of these fields, we construct a metric: the global-mean SAT, the difference in precipitation between the Northern and Southern hemispheres as a measure of the shift in hemispheric rainfall (PRE), the total sea-ice area in the Southern Hemisphere (SHI) and the temperature around the Antarctic coast at a depth of 400 m (taken as the temperature in the nearest two grid boxes from the coast; ACT). We demonstrate that the inclusion of ice-sheet meltwater reduces global atmospheric warming, shifts rainfall northwards, and increases sea-ice area and offshore, subsurface (400-m depth) Antarctic ocean temperatures. We thus show that ice-sheet meltwater should be accounted for in climate models to improve projections, because it is likely to affect policy targets.

¹Department of Geosciences, University of Arizona, Tucson, AZ, USA. ²Geophysical Fluid Dynamics Laboratory, Princeton University Forrestal Campus, Princeton, NJ, USA. ³Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ, USA. *e-mail: benjamin.bronselaer@noaa.gov

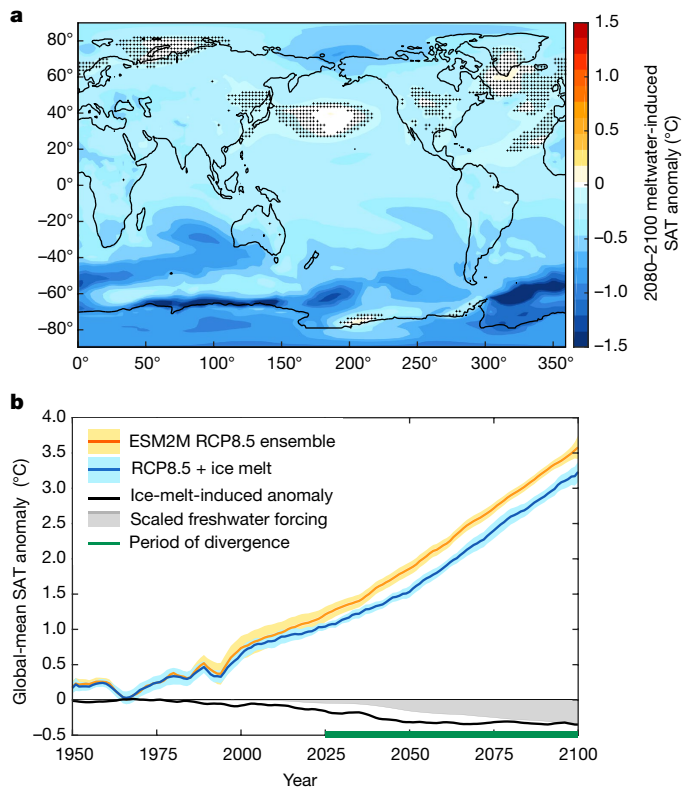


Fig. 1 | Surface air temperature anomalies. **a**, 2080–2100 meltwater-induced SAT anomaly relative to the standard ensemble (colour scale). Hatching indicates where the anomalies are not significant at the 95% level. **b**, Time series of the global-mean SAT anomaly relative to the 1950–1970 mean. Orange shows the standard ensemble and blue shows the meltwater ensemble. Solid lines show ensemble means, the dark shading shows the 95% uncertainty in the mean and the light shading shows the full ensemble spread of 20-year means. (In this case, the dark shading is too narrow to be visible.) The solid black line shows the difference between the orange and blue lines, and the applied meltwater flux is shown in grey (scaled to the mean of the final five years of the meltwater-induced SAT anomaly). The green bar indicates the period when the standard and meltwater ensembles diverge.

Surface air temperature

The release of meltwater around the Antarctic coast results in cooling of the surface ocean and overlying atmosphere relative to the RCP8.5 scenario (Fig. 1a). The largest meltwater-induced temperature anomalies are simulated throughout the Southern Hemisphere and extend into most of the Northern Hemisphere, mitigating some of the warming due to RCP8.5 greenhouse gas emissions throughout the globe.

Time evolution of the global-mean SAT shows that this meltwater-induced cooling translates to a reduced rate of global warming (Fig. 1b). The maximum difference between the two ensembles (meltwater and standard) occurs in the year 2055, when the meltwater-induced cooling is 0.38 ± 0.02 °C (95% uncertainty range).

The SAT response and the forcing curve (Fig. 1b) show that the former is not linearly related to meltwater. Rather, it becomes weaker as the ocean becomes more stratified. The ocean stratifies as a result of both warming and freshening at the surface, so ice-sheet meltwater has a weaker overall effect on stratification as the ocean surface warms and the background convection reduces (for example, in the extreme case in which there is no ocean convection, any additional surface freshening has no further effect on convection).

Precipitation

Global changes in freshwater availability are determined by rainfall that can be characterized by changes in the position of the Intertropical Convergence Zone (ITCZ). The meltwater ensemble shows a northward

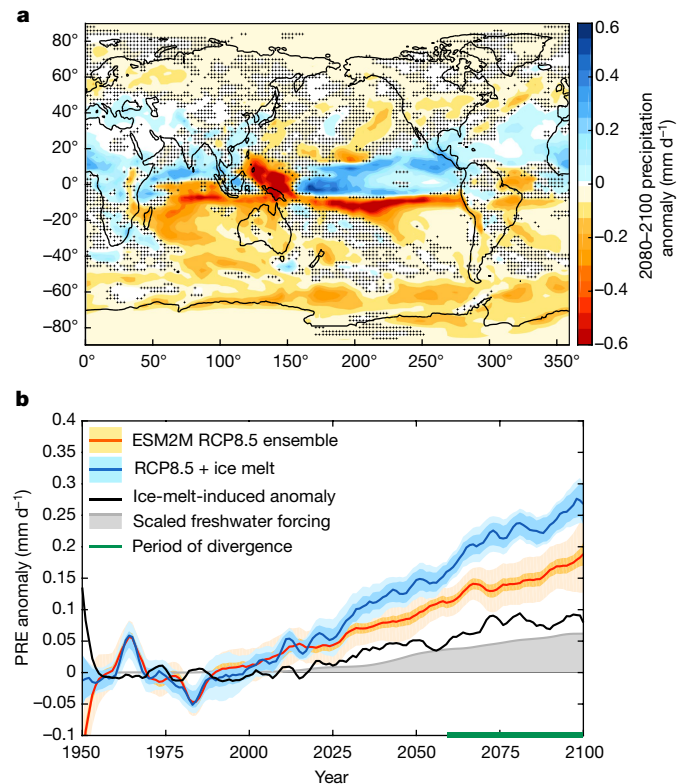


Fig. 2 | Precipitation anomalies. **a**, 2080–2100 meltwater-induced precipitation anomaly relative to the standard ensemble. Hatching indicates where the anomalies are not significant at the 95% level. **b**, Time series of the PRE anomaly relative to the 1950–1970 mean. Lines and shading as in Fig. 1.

shift of the ITCZ compared to the standard ensemble, away from the hemisphere where meltwater is added (Fig. 2a). This finding is consistent with previous work^{21,22}, which showed that additional freshwater release in the northern Atlantic Ocean causes a southward shift in the ITCZ, towards the warmer hemisphere.

The meltwater-induced precipitation change is strongest near the equator. The time evolution of the position of the ITCZ shows a gradually increasing shift towards the Northern Hemisphere in both ensembles (Fig. 2b); however, the measured shift is stronger in the meltwater ensemble. Unlike the SAT anomaly, we find that the PRE anomaly changes linearly with meltwater flux (with linear regression coefficient of determination $R^2 = 0.92$).

Although the largest changes in precipitation occur over the ocean, the changes in rainfall over land generally follow the shift in the ITCZ: increased rainfall north of the equator and decreased rainfall in the Southern Hemisphere. The ice-melt-induced precipitation changes can affect El Niño–Southern Oscillation patterns, reduce drying of semi-arid regions in the Northern Hemisphere (such as Central America; Extended Data Fig. 2) and increase drying south of the equator (for example, in Australia). Each change will have important consequences for agriculture and water scarcity.

Southern Hemisphere sea-ice area

Meltwater causes an increase in annual-mean SHI relative to the RCP8.5 scenario (Fig. 3a), which is dominated mostly by winter sea-ice anomalies (Extended Data Fig. 3). The maximum SHI anomaly occurs in the year 2055. In this year, the mean SHI anomaly is $31\% \pm 3\%$ of the pre-industrial annually averaged SHI. However, the SHI anomaly declines in the second half of the twenty-first century. After the year 2060, SHI reduces as the ocean surface continues to warm. At the end of the twenty-first century, the meltwater ensemble projects almost no change in SHI compared to the 1950–1970 mean, as opposed to a 10% reduction in the standard ensemble.

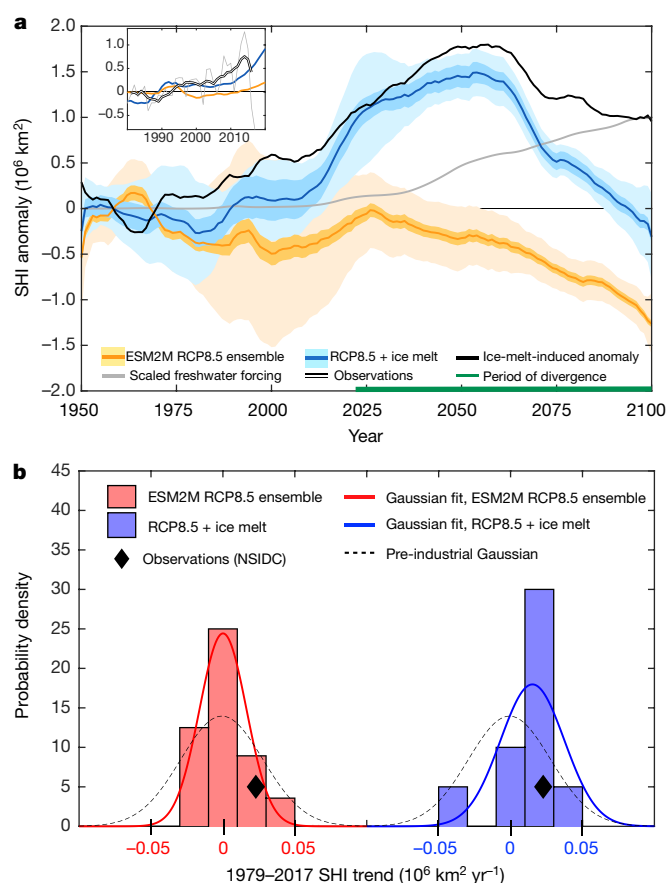


Fig. 3 | Sea-ice anomalies. **a**, Time series of the annual-mean SHI anomaly relative to the 1950–1970 mean. Lines as in Fig. 1. The inset shows the period 1980–2020, with the double black line showing the observed SHI anomaly from the National Snow and Ice Data Center (NSIDC)⁴⁶, relative to the 1980–2000 mean, and the thin grey line showing the unsmoothed observations. **b**, Distribution of linear trends in SHI over the period 1979–2017, calculated for each ensemble member. The red bars show the standard ensemble and blue bars show the meltwater ensemble, with different x axes. The solid lines show Gaussian fits to the distributions, and the dashed black line shows the pre-industrial distribution. The NSIDC observations are shown as black diamonds.

The increase in SHI in the meltwater ensemble that peaks in the year 2055 begins at the start of the twenty-first century. This increase is in contrast to the predictions of most climate model simulations, which show declining SHI²³, but is in line with the observed SHI trend over the period 1979–2017 (Fig. 3b). The reason for the observed trend is uncertain: it could be explained, wholly or partly, by natural variability^{20,24}, forced atmospheric circulation changes^{25,26} or increased freshwater input from ice shelves^{11,26,27}. Over the period 1994–2012, the additional freshwater flux in the model due to meltwater is 0.01 Sv (1 Sv = $10^6 \text{ m}^3 \text{ s}^{-1}$); the observational estimate is 0.004–0.017 Sv^{1,28–30}. The rate of change of freshwater flux in the model is 0.0007 Sv yr⁻¹ over this period, with observations estimating this flux to be 0.0007–0.0013 Sv yr⁻¹. Although the mean additional freshwater flux in the model over this period is in the middle of the observational range, the rate of change of the flux is at the low end.

The distributions of 1979–2017 linear trends in the standard and meltwater ensembles and in observations are shown in Fig. 3b. The standard ensemble has a weak mean trend ($(0.000065 \pm 0.003) \times 10^6 \text{ km}^2 \text{ yr}^{-1}$), similar to the pre-industrial (before 1850) distribution of 39-year SHI trends, as diagnosed from a 1,500-year control simulation. The distribution of trends in the meltwater ensemble has a positive mean value of $(0.015 \pm 0.007) \times 10^6 \text{ km}^2 \text{ yr}^{-1}$. Both ensembles can be considered to be consistent with observations, because the observational trend lies within the simulated range of natural variability

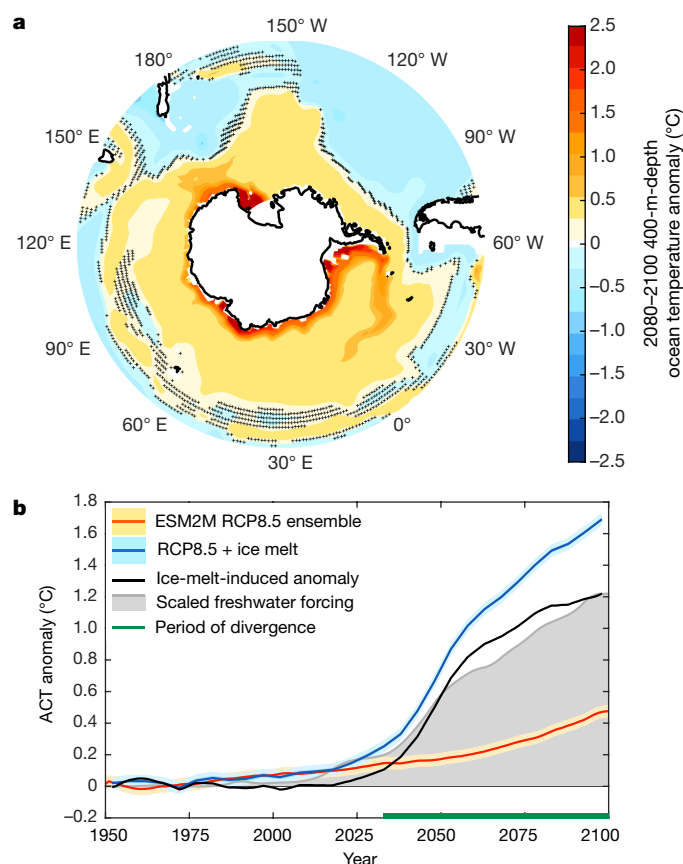


Fig. 4 | Ocean warming. **a**, 2080–2100 meltwater-induced anomaly in the ocean temperature around the Antarctic coast at a depth of 400 m, relative to the standard ensemble. Hatching indicates where the anomalies are not significant at the 95% level. **b**, Time series of the ACT anomaly relative to the 1950–1970 mean. Lines as in Fig. 1.

of each. However, we find that the observational trend lies closer to the mean of the meltwater ensemble mean than to the mean of the standard ensemble. Owing to natural variability, we cannot attribute the observed SHI trend to a flux of ice-sheet meltwater, but it is likely to contribute. The meltwater ensemble simulates a large positive trend in SHI over the entire first half of the twenty-first century, so we cannot rule out a continued increase.

Antarctic coastal warming

The meltwater-induced subsurface ocean warming around Antarctica (Fig. 4a) is highly concentrated along the coast in the Ross and Weddell seas, where it exceeds 3.5 °C and 2.5 °C, respectively. Although the time evolution of the ACT anomaly in the standard ensemble also shows warming³¹ (Fig. 4b), it is a part of a widespread pattern of warming without enhanced warming around the coast. The subsurface warming in the meltwater ensemble reaches a maximum at the end of the twenty-first century, increasing almost linearly with the strength of the meltwater flux.

Meltwater-induced ocean warming is mostly focused in the upper 1,000 m of the water column (Fig. 5a). It first appears in 1995, when it is largest at a depth of 1,250 m; however, as the atmosphere warms and the meltwater flux increases, the maximum warming increases in strength and shoals towards the surface. This is because the coastal meltwater-induced changes in stratification become more confined to the surface. The heat-flux anomalies that arise from the meltwater flux (Extended Data Fig. 4) are caused predominantly by eddy-induced isopycnal transport—that is, advection and diffusion of heat by parameterized mesoscale eddies along surfaces of constant density³². Isopycnals are depressed near the Antarctic coast, owing to surface freshening by the meltwater, causing transport of relatively warm CDW

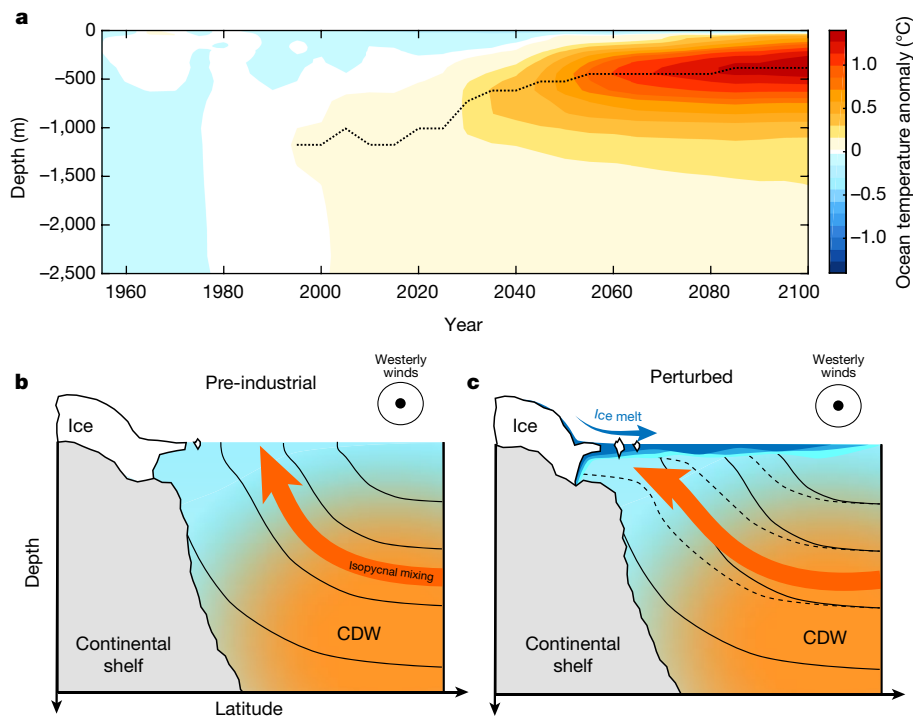


Fig. 5 | Mechanism for ocean warming. **a**, Hovmöller diagram of the meltwater-induced ocean temperature anomaly, averaged along the Antarctic coast, as a function of time. The black dotted line indicates the maximum warming in a given year. **b**, **c**, Schematic of the meltwater-induced Southern Ocean subsurface warming, shown as a zonal-mean cross-section. In the pre-industrial state (**b**), isopycnals (black lines) are tilted towards the ocean surface by westerly winds (black circles, directed out of the page), away from the continental shelf, with an upward heat

towards the Antarctic continental shelf rather than towards the surface (away from the shelf; Fig. 5b).

The discharge of meltwater causes ocean warming at the depth at which water masses are in contact with ice shelves. This warming can enable a positive feedback: warmer ocean waters increase subsurface-ice melting, which in its turn leads to more meltwater and subsequently further subsurface ocean warming^{9,33,34}. This feedback was not taken into account in a previous estimate⁴ of ice-sheet melt, which uses ocean temperatures from an uncoupled ocean model.

Implications

The climate metrics that we consider (SAT, PRE, SHI and ACT) lead to substantially different future climate projections when accounting for the effects of meltwater from the Antarctic Ice Sheet. These differences have consequences for climate policy and should be taken into account in future IPCC reports, given recent observational evidence of increasing mass loss from Antarctica³⁵. A simulated COP21 maximum global-mean atmospheric warming target of 1.5 °C relative to the pre-industrial period is first reached in the year 2037 in the standard ensemble, but in the year 2050 in the meltwater ensemble. Similarly, 2 °C of warming is first reached in 2053 in the standard ensemble, but only in 2065 in the meltwater ensemble. However, this meltwater-induced reduction in transient climate warming occurs in tandem with the potential for enhanced sea-level rise. These results emphasize the importance of the response of the Southern Ocean to ice-sheet mass loss for estimates of twenty-first-century climate change, and thus the need to account for meltwater effects in climate projections. The direct contribution from mass loss from the Antarctic Ice Sheet is already included in the IPCC assessments of future sea-level rise, although it is acknowledged to be highly uncertain in the fifth assessment report. However, the effect on climate is not included, and will not be in the

flux transporting heat from the warm CDW (orange water) towards the cooler surface (blue water), as shown by the red arrow. In the perturbed state (**c**), meltwater from the Antarctic Ice Sheet freshens the surface (blue), depressing isopycnals (solid to dashed black lines) so that isopycnal mixing transports heat towards the continent rather than towards the ocean surface (red arrow), leading to coastal warming at depth around the shelf and cooling at the surface.

upcoming CMIP6 experimental design. Similarly, the effects of meltwater from the Greenland Ice Sheet have so far not been considered, and could lead to further changes in simulated future climate^{8,36}.

We identify a physical mechanism whereby increased meltwater could lead to heat transport, enhanced by eddies, into cavities beneath the Antarctic ice shelf, enabling a positive feedback. Although our model does not resolve ice-shelf cavities, we can estimate the future magnitude of this positive feedback on ice-shelf melt using an existing parameterization⁴ (Methods). Meltwater-induced ocean warming could result in an increase in Southern Ocean meltwater flux of 9%–34% as a result of increased ice-shelf melt (Extended Data Fig. 5), even when we consider the possibility of lower meltwater flux. This feedback could potentially increase sea level further by causing an increased ice flux across the grounding line of the ice sheet. Such a feedback mechanism is supported by palaeo evidence^{19,37}. However, the Southern Ocean is a complex system and many ice-sheet-related feedbacks need to be accounted for, such as atmospheric heat and moisture transport, surface heat fluxes, ice-shelf-cavity dynamics and sea-ice changes^{22,35,38}. Global coupled models such as ESM2M are useful tools for identifying and quantifying the potential of this feedback by modelling the temperature response of the Southern Ocean to meltwater discharge and accounting for global feedbacks that process-based models cannot capture; however, they lack high-resolution continental-shelf and ice-shelf-cavity dynamics. Consequently, global coupled model simulations need to be complemented with regional ice-sheet studies to fully constrain the magnitude of the ice-loss feedback. The most recent model estimate⁴ of the Antarctic contribution to global sea level in 2100 is 0.86 m; but, considering feedbacks such as meltwater-induced ocean warming, this number could be higher.

Although here we mainly discuss the changes in the ensemble-mean climate state when including ice-sheet meltwater, the time series in Figs. 1b, 2b, 3a and 4b show the period during which the meltwater

signal would be significant in a single climate simulation, such as a single submission to CMIP6. For the SAT, SHI and ACT metrics, this period is most of the twenty-first century, which demonstrates the distinct effect on climate of ice-sheet melt over natural variability and the importance of including the associated meltwater flux in all simulations. Coupling ice-sheet models to climate models remains a challenge, but as a feasible intermediate step we recommend adding projected meltwater flux, although it is not mass-conserving.

There are several caveats in our experimental design. We impose the meltwater flux in a spatially uniform pattern around the coast at the surface of the ocean, without partitioning into liquid meltwater and solid icebergs. A previous study shows that most of the meltwater input from icebergs occurs within our meltwater flux region around the Antarctic coast³⁹, which justifies neglecting iceberg meltwater injection. Ice-sheet mass loss is also not uniform^{1,40}. However, it is unclear whether the spatially uniform injection of meltwater affects the effect of meltwater on climate^{41,42} (Methods). We add the meltwater at the surface even though some will be discharged at depth owing to basal melt. Doing so might affect regional sea-ice trends²⁶, but does not greatly affect the trends in SHI and sea-surface temperature³⁰. Further research with coupled ice–ocean–atmosphere models should therefore focus on constraining the effect of meltwater-induced subsurface ocean warming on ice-shelf melt. Although observational sea-ice trends are highly regional and probably depend on the spatial distribution of the sea-ice melt²⁶, our uniform flux is appropriate for analysing SHI³⁰. In addition, we use only one climate model, and we expect the quantitative results to be model-dependent, including the estimates of the effect of the meltwater feedback mechanism^{8,43}. However, the large-scale ocean and atmospheric mechanisms that lead to the SAT, PRE, SHI and ACT anomalies shown here in response to Antarctic meltwater discharge should be robust because they are consistent with previous studies^{8,9,18,19,44}. The multi-model response should be assessed through efforts such as the Southern Ocean Modelling Intercomparison Project (SOMIP; <http://southernocean.arizona.edu/node/42>) and the Flux-Anomaly-Forced Model Intercomparison Project (FAFMIP)⁴⁵.

Conclusions

Our study, which focused on accounting for the effects of Antarctic Ice Sheet meltwater on the rest of the climate system using large ensemble RCP8.5 climate simulations, finds that these effects are substantial and that meltwater discharge is important when determining the state of the climate in the twenty-first century. Meltwater causes a reduction in global atmospheric warming, delaying the realization of 1.5 °C and 2 °C warming by more than ten years; it drives a northward shift of the ITCZ, which results in reduced drying over Northern Hemisphere landmasses and enhanced drying in the Southern Hemisphere; and it causes a large (up to 31%) increase in Antarctic sea-ice formation relative to the pre-industrial period and an increase in subsurface ocean warming around the Antarctic coast by a factor of four. Our results suggest that a feedback mechanism is in operation, whereby the meltwater-induced subsurface warming leads to enhanced melting underneath ice shelves, potentially causing further meltwater-related climate effects. Our results demonstrate that meltwater discharge from the Antarctic Ice Sheet not only contributes to sea-level rise but also influences the global climate throughout most of the twenty-first century, emphasizing the importance of ocean and ice-sheet feedbacks on the climate system. Antarctic meltwater is therefore an important agent of climate change with global impact, and should be taken into account in future climate simulations and climate policy.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0712-z>.

Received: 9 April; Accepted: 5 September 2018;
Published online 19 November 2018.

- Paolo, F. S., Fricker, H. A. & Padman, L. Volume loss from Antarctic ice shelves is accelerating. *Science* **348**, 327–331 (2015).
- Wouters, B. et al. Dynamic thinning of glaciers on the southern Antarctic peninsula. *Science* **348**, 899–903 (2015).
- Konrad, H. et al. Net retreat of Antarctic glacier grounding lines. *Nat. Geosci.* **11**, 258–262 (2018).
- DeConto, R. M. & Pollard, D. Contribution of Antarctica to past and future sea-level rise. *Nature* **531**, 591–597 (2016).
- Taylor, K. E., Stouffer, R. J. & Meehl, G. A. An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* **93**, 485–498 (2012).
- Eyring, V. et al. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.* **9**, 1937–1958 (2016).
- Rignot, E., Velicogna, I., van den Broeke, M. R., Monaghan, A. & Lenaerts, J. Acceleration of the contribution of the Greenland and Antarctic ice sheets to sea level rise. *Geophys. Res. Lett.* **38**, L05503 (2011).
- Stouffer, R. J., Seidov, D. & Haupt, B. J. Climate response to external sources of freshwater: North Atlantic versus the Southern Ocean. *J. Clim.* **20**, 436–448 (2007).
- Fogwill, C. J., Phipps, S. J., Turney, C. S. M. & Golledge, N. R. Sensitivity of the Southern Ocean to enhanced regional Antarctic ice sheet meltwater input. *Earth's Futur.* **3**, 317–329 (2015).
- Park, W. & Latif, M. Ensemble global warming simulations with idealized Antarctic meltwater. *Clim. Dyn.* <https://doi.org/10.1007/s00382-018-4319-8> (2018).
- Bintanja, R., van Oldenborgh, G. J., Drijfhout, S. S., Wouters, B. & Katsman, C. A. Important role for ocean warming and increased ice-shelf melt in Antarctic sea-ice expansion. *Nat. Geosci.* **6**, 376–379 (2013).
- Pauling, A. G., Smith, I. J., Langhorne, P. J. & Bitz, C. M. Time-dependent freshwater input from ice shelves: impacts on Antarctic sea ice and the Southern Ocean in an Earth system model. *Geophys. Res. Lett.* **44**, 10454–10461 (2017).
- Rhodes, C. J. The 2015 Paris climate change conference: COP21. *Sci. Prog.* **99**, 97–104 (2016).
- Oppenheimer, M. Global warming and the stability of the West Antarctic Ice Sheet. *Nature* **393**, 325–332 (1998).
- Rignot, E. & Jacobs, S. Rapid bottom melting widespread near Antarctic ice sheet grounding lines. *Science* **296**, 2020–2023 (2002).
- Shepherd, A., Wingham, D. & Rignot, E. Warm ocean is eroding West Antarctic Ice Sheet. *Geophys. Res. Lett.* **31**, L23402 (2004).
- Obase, T., Abe-Ouchi, A., Kusahara, K., Hasumi, H. & Ohgaito, R. Responses of basal melting of Antarctic ice shelves to the climatic forcing of the Last Glacial Maximum and CO₂ doubling. *J. Clim.* **30**, 3473–3497 (2017).
- Aiken, C. M. & England, M. H. Sensitivity of the present-day climate to freshwater forcing associated with Antarctic sea ice loss. *J. Clim.* **21**, 3936–3946 (2008).
- Bakker, P., Clark, P. U., Golledge, N. R., Schmittner, A. & Weber, M. E. Centennial-scale Holocene climate variations amplified by Antarctic Ice Sheet discharge. *Nature* **541**, 72–76 (2017).
- Swart, N. C. & Fyfe, J. C. The influence of recent Antarctic ice sheet retreat on simulated sea ice area trends. *Geophys. Res. Lett.* **40**, 4328–4332 (2013).
- Zhang, R. & Delworth, T. Simulated tropical response to a substantial weakening of the Atlantic thermohaline circulation. *J. Clim.* **18**, 1853–1860 (2005).
- Cabr  , A., Marinov, I. & Gnanadesikan, A. Global atmospheric teleconnections and multidecadal climate oscillations driven by Southern Ocean convection. *J. Clim.* **30**, 8107–8126 (2017).
- Purich, A., Cai, W., England, M. H. & Cowan, T. Evidence for link between modelled trends in Antarctic sea ice and underestimated westerly wind changes. *Nat. Commun.* **7**, 10409 (2016).
- Polvani, L. M. & Smith, K. L. Can natural variability explain observed Antarctic sea ice trends? New modeling evidence from CMIP5. *Geophys. Res. Lett.* **40**, 3195–3199 (2013).
- Haumann, F. A., Notz, D. & Schmidt, H. Anthropogenic influence on recent circulation-driven Antarctic sea ice changes. *Geophys. Res. Lett.* **41**, 8429–8437 (2014).
- Merino, N. et al. Impact of increasing antarctic glacial freshwater release on regional sea-ice cover in the Southern Ocean. *Ocean Model.* **121**, 76–89 (2018).
- Bintanja, R., Van Oldenborgh, G. J. & Katsman, C. A. The effect of increased fresh water from Antarctic ice shelves on future trends in Antarctic sea ice. *Ann. Glaciol.* **56**, 120–126 (2015).
- Shepherd, A. et al. A reconciled estimate of ice-sheet mass balance. *Science* **338**, 1183–1189 (2012).
- Sutterley, T. C. et al. Mass loss of the Amundsen sea embayment of West Antarctica from four independent techniques. *Geophys. Res. Lett.* **41**, 8421–8428 (2014).
- Pauling, A. G., Bitz, C. M., Smith, I. J. & Langhorne, P. J. The response of the Southern Ocean and Antarctic sea ice to freshwater from ice shelves in an Earth system model. *J. Clim.* **29**, 1655–1672 (2016).
- Goddard, P. B., Dufour, C. O., Yin, J., Griffies, S. M. & Winton, M. CO₂-induced ocean warming of the Antarctic continental shelf in an eddy global climate model. *J. Geophys. Res. Oceans* **122**, 8079–8101 (2017).
- Stewart, A. L. & Thompson, A. F. Eddy-mediated transport of warm circumpolar deep water across the Antarctic shelf break. *Geophys. Res. Lett.* **42**, 432–440 (2015).
- Silvano, A. et al. Freshening by glacial meltwater enhances melting of ice shelves and reduces formation of Antarctic bottom water. *Sci. Adv.* **4**, eaap9467 (2018).

34. Spence, P. et al. Localized rapid warming of West Antarctic subsurface waters by remote winds. *Nat. Clim. Chang.* **7**, 595–603 (2017).
35. Massom, R. A. et al. Antarctic ice shelf disintegration triggered by sea ice loss and ocean swell. *Nature* **558**, 383–389 (2018).
36. Vizzaino, M. et al. Coupled simulations of Greenland Ice Sheet and climate change up to AD 2300. *Geophys. Res. Lett.* **42**, 3927–3935 (2015).
37. Sangiorgi, F. et al. Southern Ocean warming and Wilkes Land ice sheet retreat during the mid-Miocene. *Nat. Commun.* **9**, 317 (2018).
38. Fyke, J., Sergeenko, O., Loftverstorm, M., Price, S. & Lenaerts, J. T. M. An Overview of Interactions and Feedbacks Between Ice Sheets and the Earth System. *Rev. Geophys.* **56**, 361–408 (2018).
39. Stern, A. A., Adcroft, A. & Sergienko, O. The effects of Antarctic iceberg calving-size distribution in a global climate model. *J. Geophys. Res. Oceans* **121**, 5773–5788 (2016).
40. Rignot, E., Jacobs, S., Mouginot, J. & Scheuchl, B. Ice-shelf melting around Antarctica. *Science* **341**, 266–270 (2013).
41. Stammer, D. Response of the global ocean to Greenland and Antarctic ice melting. *J. Geophys. Res. Oceans* **113**, C06022 (2008).
42. Haid, V., Iovino, D. & Masina, S. Impacts of freshwater changes on Antarctic sea ice in an eddy-permitting sea-ice-ocean model. *Cryosphere* **11**, 1387–1402 (2017).
43. He, J., Winton, M., Vecchi, G., Jia, L. & Rugenstein, M. Transient climate sensitivity depends on base climate ocean circulation. *J. Clim.* **30**, 1493–1504 (2017).
44. Swingedouw, D., Fichefet, T., Goosse, H. & Loutre, M. F. Impact of transient freshwater releases in the Southern Ocean on the AMOC and climate. *Clim. Dyn.* **33**, 365–381 (2009).
45. Gregory, J. M. et al. The Flux-Anomaly-Forced Model Intercomparison Project (FAFMIP) contribution to CMIP6: investigation of sea-level and ocean climate change in response to CO₂ forcing. *Geosci. Model Dev.* **9**, 3993–4017 (2016).
46. Fetterer, F., Knowles, K., Meier, W., Savoie, M. & Windnagel, A. K. Sea ice index, version 3: sea ice extent. *National Snow and Ice Data Center* <https://nsidc.org/data/G02135/versions/3> (2017).
73. NOAA. *Data Announcement 88-MGG-02, Digital Relief of the Surface of the Earth* <https://www.ngdc.noaa.gov/mgg/global/etopo5.HTML> (National Geophysical Data Center, Boulder, 1988).

Acknowledgements This work was supported by the NSF's Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) Project under the NSF award PLR-1425989, the NOAA and NASA. O.V.S. is supported by NSF OPP-1246151 and by NOAA awards NA14OAR4320106 and NA13OAR431009 from the US Department of Commerce. Support for K.B.R. was provided by NASA award NNX17AI75G. The statements, findings, conclusions and recommendations presented here are those of the authors and do not necessarily reflect the views of the NOAA or the US Department of Commerce. We thank J. Sarmiento, J. Yin and A. Haumann for their insight.

Author contributions B.B. performed the simulations with help from W.J.H., M.W., R.J.S. and K.B.R.; B.B., M.W. and J.L.R. analysed the data with help from R.J.S. (climate response to ocean meltwater input), S.M.G. (ocean dynamics and heat budget) and O.V.S. (ice–ocean interactions and feedbacks); M.W. and J.L.R. supervised the project. All authors wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0712-z>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to B.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Projecting ice-melt freshwater flux. To represent Antarctic Ice Sheet meltwater, we apply an external source of freshwater to the climate model. We obtain a time series of yearly freshwater flux representing melt from the ice sheets and shelves of Antarctica by digitizing extended data figure 8 of ref. ⁴. The error in total applied freshwater flux arising from the digitization in the year 2100 is 1.3% (Extended Data Fig. 1), which is negligible in the context of our study (the difference in the response of ACT to 200% and 50% of the flux from ref. ⁴ is only roughly $\pm 25\%$, as discussed in Methods section ‘Estimating subsurface warming feedback’). This flux includes the freshwater input from melting of the grounded ice sheet and of the floating ice shelves. It also includes contributions from basal melt, grounding-line retreat, surface meltwater and rain runoff, surface and basal calving, crevassing and hydrofracturing. The freshwater flux gives the total change in ice-sheet volume, which accounts for the accumulation of precipitation over the Antarctic continent as the climate warms. This change in precipitation is also simulated in ESM2M, resulting in a double counting. However, the ensemble-mean simulated cumulative precipitation change over Antarctica in 2100 in the standard ensemble is 1% of the total freshwater flux and therefore negligible.

Throughout this paper, we refer to both ice-sheet and ice-shelf melt. The ice-sheet melt refers to meltwater from grounded ice and ice-shelf melt refers to meltwater from floating ice. Although it is the total freshwater flux that is relevant for the climate response to ice melt, it is only the ice-sheet melt that will contribute to global-mean sea-level rise (roughly half of the total freshwater input). The equivalent global-mean sea-level rise for the prescribed freshwater flux is therefore higher than the global-mean sea-level rise quoted in ref. ⁴. The ice-sheet model used in ref. ⁴ uses a 10-km horizontal resolution coupled to a regional atmospheric model. Ocean temperatures for basal melt in ref. ⁴ were taken from the 400-m depth level from a de-coupled RCP8.5 simulation performed with the NCAR CCSM4⁴⁷. The model was forced with total equivalent atmospheric CO₂ accounting for the contributions from radiatively active trace gases.

GFDL ESM2M. The model used in this study is GFDL ESM2M. ESM2M is a CMIP5 Earth system model with a full carbon cycle at $1^\circ \times 1^\circ$ horizontal ocean model resolution with increased resolution near the equator and 50 unevenly spaced vertical levels in depth coordinates, with a free surface^{48,49} and mesoscale eddies parameterized using the GM-Redi schemes⁵⁰. ESM2M does not have interactive ice sheets, so ice-sheet and shelf-melt need to be prescribed as freshwater flux. The transient climate response (TCR) of ESM2M is 1.5 K, which is at the low end of the CMIP5 models⁵¹. The Southern Ocean in ESM2M has relatively deep mixed layers compared to the overall CMIP5 mean, resulting from a slightly more convective mean state relative to the CMIP5 mean, but it is not an outlier within the CMIP5 ensemble⁵². The ESM2M mean state could affect the magnitude of the response of the model to freshwater flux compared to other models⁴³.

The response of the mean depth of the Southern Ocean mixed layer to future climate change scenarios (RCP4.5 and RCP8.5) is weak compared to the CMIP5 mean⁵². The mean-state Antarctic SHI is low compared to observations⁵³. The ESM2M negative historical (1979–2013) trend in Antarctic summer and winter sea-ice extent and Southern Ocean sea-surface temperature warming south of 55° S is weaker than average²³. ESM2M simulates summertime cooling of Southern Ocean sea-surface temperatures over the period 1979–2013, whereas most other models simulate warming²³. Although ESM2M is on the convective side of CMIP5 models, the sea-ice volume varies only weakly with the rate of Southern Ocean convection⁵⁴. It has been shown⁵² that there is a large spread in CMIP5-simulated winter-time mixed-layer patterns and convection. However, ESM2M has deep mixed layers in areas that correspond to the observations. Winter-time open-ocean convection and polynya formation in ESM2M is confined to the eastern Weddell Sea⁵², similar to observed open-ocean polynyas in the 1970s and in recent years^{55,56}. Observations also show deep mixed layers around the Antarctic coast where dense water is thought to form. ESM2M also simulates deeper mixed layers around the coast; however, these mixed layers are shallower than observed⁵⁷. CMIP5 models are also known to have Southern Ocean westerly winds that are too far equatorwards⁵⁸. However, ESM2M has one of the smallest biases in Southern Ocean wind position and strength among CMIP5 models⁵⁹. Moreover, ESM2M has been shown to perform well in terms of global and Southern Ocean heat uptake⁶⁰. We compare the density structure of ESM2M around the Antarctic coast to the Southern Ocean state estimate (SOSE)⁶¹ in Methods section ‘ESM2M Southern Ocean evaluation’.

Experimental design. To simulate melting of Antarctic land ice, we add an external source of freshwater at the surface of the ocean to ten ensemble members, similarly to previous studies⁸. The additional freshwater is added at sea-surface temperature. The freshwater perturbation does not have a seasonal cycle and is added uniformly around the Antarctic continent, in the three grid boxes against the coast (corresponding to 3° in latitude away from the coastline). Although we impose all of the freshwater at the surface of the ocean for ease of reproduction, part of this melt is due to basal melting, which would be deposited in the ocean at a depth of several hundred metres. It has been shown³⁰ that putting all of the

freshwater flux at depth compared to all at the surface affects the response of the local mixed layer, but does not greatly affect SHI trends. Adding the freshwater at depth increases the magnitude of subsurface warming and the depth of the freshwater flux affects regional sea-ice trends³⁰. For these reasons, we acknowledge that adding all of the freshwater at the surface is a limitation, but one that is unlikely to substantially affect the climate metrics discussed here, apart from the subsurface warming, which may be underestimated.

We also add the freshwater uniformly around the Antarctic coast for ease of reproduction and because climate models simulate different stratification spatial patterns around Antarctica⁵². A uniform flux would therefore make results more comparable when we assess the multi-model effect of the freshwater flux for use in IPCC projections. The effect of spatially varying meltwater flux on the climate compared to a uniform flux is beyond the scope of our study.

The ten freshwater perturbation experiments branch from a randomly selected subset of the initial-condition perturbations on 1 January 1950 that are used for the 30 ESM2M large ensemble simulations⁶² (Extended Data Fig. 6). These new freshwater perturbation experiments follow the same historical and RCP8.5 boundary conditions⁵ that were used for the ESM2M large ensemble runs, differing only from the earlier run in their added freshwater perturbation. As stated above, the magnitude of the freshwater perturbation follows a previous projection of Antarctica ice-sheet melt⁴. This set of simulations therefore shows directly the effect on the climate of adding projected ice-sheet and ice-shelf melt to the relevant climate change scenario.

It is important that the response of the climate to ice-sheet melt is assessed in a fully coupled global model in a climate change scenario, because the sensitivity of the system to freshwater flux will change as the climate warms⁴³ (Extended Data Fig. 7). There are feedbacks in the system that can be quantified only with global coupled climate models. Southern Ocean freshwater affects large-scale ocean dynamics and ocean–atmospheric coupling, which in turn have been shown to influence Southern Ocean stratification, convection and sea ice on multiple timescales^{63,64}. For example, it has been shown²² that a shift in the tropical Hadley circulation influences Southern Ocean stratification and convection. As shown in Fig. 2, Antarctic Ice Sheet melt results in a persistent northward shift of the ITCZ and Hadley circulation, which will feedback into Southern Ocean stratification. Dynamic sea ice has also been shown to influence subsurface ocean temperatures. Ocean stratification is key for correctly simulating the time-varying sensitivity (Extended Data Figs. 4, 7) of freshwater-forced subsurface coastal warming because the main driving terms are isopycnal mixing and stirring. Such processes are captured by well-tested CMIP5-class models such as ESM2M. Our results are relevant for CMIP5 and CMIP6 projections, so it is important that these results are demonstrated in a CMIP5 model such as ESM2M.

Significance testing. For significance testing of ensemble-mean differences, we test for a 95% confidence level using a pair-wise test. We take the difference between each member of the meltwater ensemble and the corresponding member of the standard ensemble from which the meltwater-ensemble member branches. The anomalies in the ten-member mean are significant if the mean anomaly is larger than $1.699\sigma/\sqrt{10}$, where σ is the estimated standard deviation of the ensemble mean.

Time of divergence of ensembles. The period when the ensembles diverge, indicated by the green bar in Figs. 1b, 2b, 3a and 4b, is that when the two ensembles are statistically different at the 95% confidence level. The standard deviation in this calculation is diagnosed from successive 20-year means of each metric from the respective ensembles. Therefore, the time of divergence does not depend explicitly on the number of ensemble members. Any 20-year mean taken from a single member of the meltwater ensemble during this period is more than 95% likely to be different from the standard ensemble.

Estimating subsurface warming feedback. To estimate the increase in ice-shelf melt caused by the freshwater-induced subsurface ocean warming around Antarctica, we follow previous methodology⁴. We use the same parameterization to express ice-shelf melt rates (OM; in metres per year) as a function of the ocean temperature around the Antarctic coast at a depth of 400 m (T_O , equivalent to ACT):

$$OM = \frac{K_T \rho_w C_w}{\rho_i L_f} |T_O - T_f| (T_O - T_f)$$

where T_f is the freezing-point temperature at the base of the ice shelf and $(K_T \rho_w C_w)/(\rho_i L_f) = 0.224 \text{ m yr}^{-1} \text{ } ^\circ\text{C}^{-2}$ (see ref. ⁴ for details of these physical parameters). To convert OM to freshwater flux, we scale it by the surface area of the ice shelf to be consistent with the data in table 1 in ref. ⁶⁵. We do so using T_O from the standard and meltwater ensembles. The difference in freshwater flux between the two ensembles gives the increase in freshwater flux that results from the ice-melt-induced subsurface warming. This freshwater flux does not include the freshwater flux across the grounding line, which is the only flux that contributes to increased sea level. However, increased shelf melt is likely to lead to an increase in grounding-line flux due to a reduction in buttressing⁶⁶.

Over the period 1995–2009, this calculation yields a basal melt-rate anomaly that is $40\% \pm 10\%$ of the total melt rate, including calving, consistent with recent observational estimates⁶⁵ of $52\% \pm 14\%$, albeit at the low end. The total basal melt over the period 1995–2009 is $1,677 \pm 771$ Gt yr⁻¹, which agrees roughly with the range of observational estimates ($1,325 \pm 235$ Gt yr⁻¹ (ref. ⁴⁰) and $1,454 \pm 174$ Gt yr⁻¹ (ref. ⁶⁵)). The resulting total cumulative freshwater flux for these calculations, expressed as equivalent global-mean sea-level rise, is shown in Extended Data Fig. 7.

We stress that the calculation shown in Extended Data Fig. 5 is a rough estimate and that a coupled ice-sheet–ocean–atmosphere model simulation is needed to fully assess the magnitude of the subsurface warming feedback. Although our simulations do not resolve ice-shelf-cavity dynamics, global atmosphere–ocean feedbacks are necessary to simulate the sensitivity of the subsurface temperature to freshwater flux (as shown by Extended Data Fig. 7), which only a comprehensive climate model such as ESM2M can capture. Our calculation assumes a constant ice-shelf area and that the total freshwater flux depends linearly on OM, and does not account for coastal and ice-shelf-cavity dynamics.

For studies that require detailed knowledge of the spatial patterns of melting below ice shelves, other parameterizations are more appropriate⁶⁷. However, the goal of our study is to quantify the net Antarctic sensitivity of ice-shelf melt to freshwater-induced ocean warming. For such integrated quantities, detailed knowledge of the spatial distribution of melting is less important. Numerous modelling studies that have focused on the sensitivity of area-averaged melt rates to ocean temperature^{68–71} have demonstrated that our parameterization adequately captures the dependence of melt rate on temperature. It is this dependence that drives the area-integrated feedback, so for our purpose the parameterization is appropriate.

To account for the effects of coastal and ice-shelf dynamics, in ref. ⁴ OM is multiplied by a dimensionless constant. For our uncertainty range, we run additional simulations in which we apply half and double the prescribed⁴ flux to the transient RCP8.5 simulation for three ensemble members. The subsurface warming in each case is shown in Extended Data Fig. 9. Doubling the freshwater flux causes an increase of roughly 28% in the freshwater-induced subsurface warming anomaly; halving it causes a 25% reduction. On the lower end, despite the freshwater flux being halved, the freshwater-induced subsurface warming is still much larger than in the standard ensemble. We then apply this transient simulated relative increase (decrease) in warming compared to the main projected⁴ freshwater flux scenario as an upper (lower) bound for ACT anomalies in the feedback calculation. We use a range of freezing point temperatures ($-1.8^\circ\text{C} < T_f < -2.6^\circ\text{C}$) and account for the uncertainty in conversion from ice-shelf melt rate in metres per year to freshwater flux using the numbers in ref. ⁴. Despite the uncertainty range in our estimate, the freshwater-forced ACT is substantially different from the standard ensemble. This difference is due to the overall large magnitude of the projected ACT anomaly, which could only be captured in a global climate model.

Seasonal sea-ice anomalies. Extended Data Fig. 3a, b shows the monthly mean SHI anomalies for February and September. Similar to the annual-mean SHI, the maximum September SHI anomaly is simulated mid-century; the February maximum is simulated around the year 2025. The maximum September-mean SHI anomaly is $24\% \pm 3\%$ of the pre-industrial SHI and the maximum February-mean SHI anomaly is $117\% \pm 20\%$ of the pre-industrial February mean.

ESM2M predicts a generally low total austral-summer sea-ice area (around 0.19×10^6 km²), which is the reason for the weak 39-year trends compared to the observational trends (Extended Data Fig. 3c, d). However, for both the February and September SHI trends, the meltwater ensemble is more consistent with the observational trends.

Sensitivity to base state. The response of the climate to ice-sheet melt should be assessed in a climate change scenario because the sensitivity to the freshwater flux will vary as the climate changes. For this reason, we add the previous⁴ estimate to a full RCP8.5 scenario⁴. To demonstrate the importance of the base state, we ran two additional 50-year ensembles with ESM2M, one initiated in the year 1980 and the other in the year 2050, each forced with a constant 0.1-Sv freshwater flux around the Antarctic coast. Each of these five-member ensembles therefore experiences the same freshwater flux for the same duration, but over a different period in the standard scenario. Extended Data Fig. 7 shows the time evolution of the SAT, PRE, SHI and ACT metrics in these ensembles compared to the standard ensemble. The magnitude of the freshwater-induced anomaly is different in each period. The SAT and SHI metrics show a reduced sensitivity in warmer climates, whereas ACT shows a larger sensitivity to the freshwater flux. The overall time-dependent anomalies for each of these metrics therefore depends on both the time-varying freshwater flux and sensitivity to the base state.

Heat budget. The heat-flux diagnostics (Extended Data Fig. 4) that we used are those described in ref. ⁷². The heat-flux term named ‘submesoscale’ refers to advection by parameterized sub-grid-scale eddies⁷². ‘Overflow’ refers to the heat transport by along-topography overflow parameterizations. ‘Vertical mixing’ includes heat fluxes due to convective mixing and to vertical diffusion. ‘Isopycnal transport’

includes parameterized eddy diffusion and eddy-induced advection by mesoscale eddies transporting tracers along surfaces of constant density. The heat-flux anomalies were diagnosed in the five-member ensemble simulation with a constant freshwater flux of 0.1 Sv over the period 1980–2030. Over this period, we find that the dominant contribution to the freshwater-induced coastal warming averaged between 400 m and 700 m is the isopycnal transport term (Extended Data Fig. 4c), which is largely due to the eddy-induced advection term.

Simulated polynyas. Five of the 30 ensemble members of the standard ensemble simulate open-ocean polynyas in the period 1970–2020. For the ten members used to branch off the freshwater simulations, we chose two of the polynya members and eight of the non-polynya members to represent the ensemble as a whole (Extended Data Fig. 6).

We tested whether there is a difference in the response of the simulations derived from polynya members versus the non-polynya members and found no considerable difference. For example, we examined the 1979–2017 sea-ice trends discussed in main-text section ‘Southern Hemisphere sea-ice area’ (Fig. 3b), which are likely to be most affected by the simulated polynyas (as shown by the spread in sea-ice extent over this period in Extended Data Fig. 6). The linear trends in the two freshwater simulations derived from polynya members are 0.029×10^6 km² yr⁻¹ and -0.0087×10^6 km² yr⁻¹, indicating no substantial bias because they lie above and below the mean of the distribution (0.015×10^6 km² yr⁻¹). There is one freshwater-perturbed member that simulates several years of convection in this period, resulting in a large negative trend in SHI of -0.042×10^6 km² yr⁻¹ (Fig. 3b). However, this is not a simulation derived from an ensemble member that simulates a polynya in the unperturbed RCP8.5 scenario. Therefore, the presence of polynyas in some members of the standard ensemble does not affect the response.

The presence of long-term polynyas simulated over the historical period in five of the 30 standard-ensemble members is not supported by observations. However, none of the meltwater-ensemble members simulates such long-term events. It therefore seems that ESM2M (and probably CMIP5 models as a whole) is more prone to simulating large convective events compared to observations, perhaps because it lacks the appropriate freshwater flux in the unperturbed simulations.

ESM2M Southern Ocean evaluation. We also compare the Southern Ocean density structure around the Antarctic coast in ESM2M with the 2008–2012 SOSE (iteration 105)⁶¹. The density structure is important for simulating the appropriate subsurface warming response to the meltwater flux. SOSE is a state estimate based on the MITgcm, constrained using available data. In particular, SOSE uses Argo and seal data to constrain the solution near the Antarctic coast. Although SOSE is a model simulation, it provides an estimate of the Southern Ocean that is consistent with available observations.

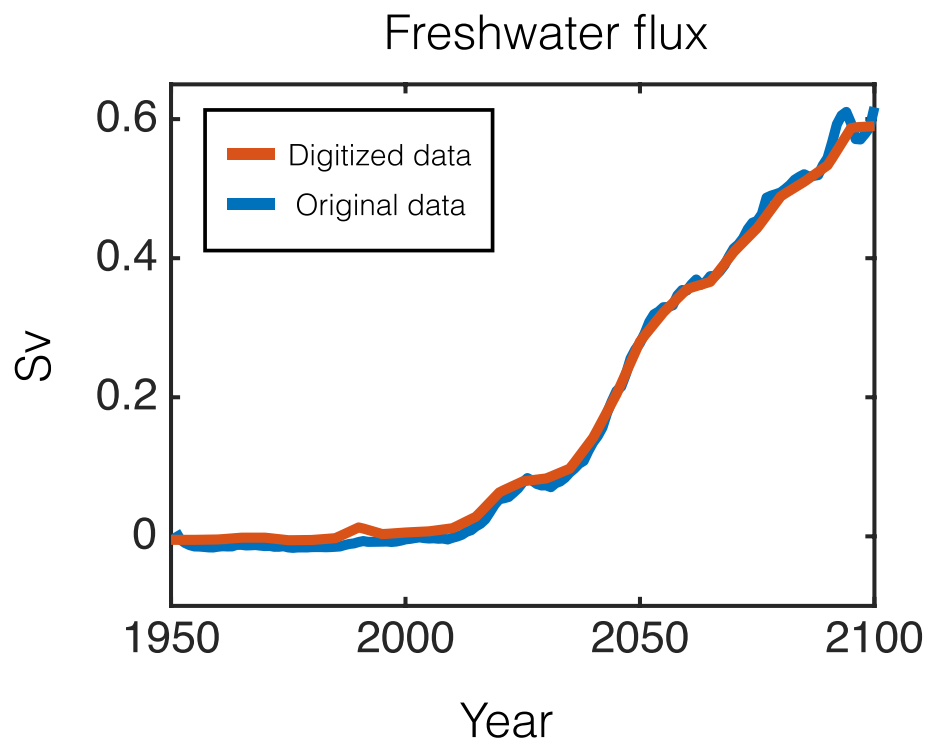
Extended Data Fig. 10a shows the ESM2M mean-state mixed-layer depth, the pattern of which agrees roughly with estimates from observations⁵⁷. Extended Data Fig. 10b shows vertical density profiles for ESM2M (black) and SOSE (red) over three key regions in which meltwater-induced subsurface warming is simulated: (1) the Ross Sea, (2) the Weddell Sea and (3) eastern Antarctica. These profiles show that the overall density structure and stratification around the Antarctic coast in ESM2M is similar to the data-constrained SOSE. Although ESM2M has a small overall bias, the vertical stratification is similar to SOSE. Extended Data Fig. 10c shows Southern Ocean zonal-mean ESM2M and SOSE isopycnal surfaces (solid and dashed, respectively), as well as local isopycnal surfaces in regions (1)–(3). South of 60° S, the ESM2M and SOSE density structures are very similar, indicating that the meltwater-induced temperature anomaly should be reasonably well represented in our simulations. However, ESM2M has a weaker isopycnal slope away from the coast compared to SOSE. This difference probably means that the meltwater-induced temperature anomaly is less confined to the coast in ESM2M than it would be given the SOSE density structure. Although this could lead to an underestimate of the warming, the figures show that the warming is mostly confined around the coast where isopycnals are mostly similar to SOSE.

Data availability

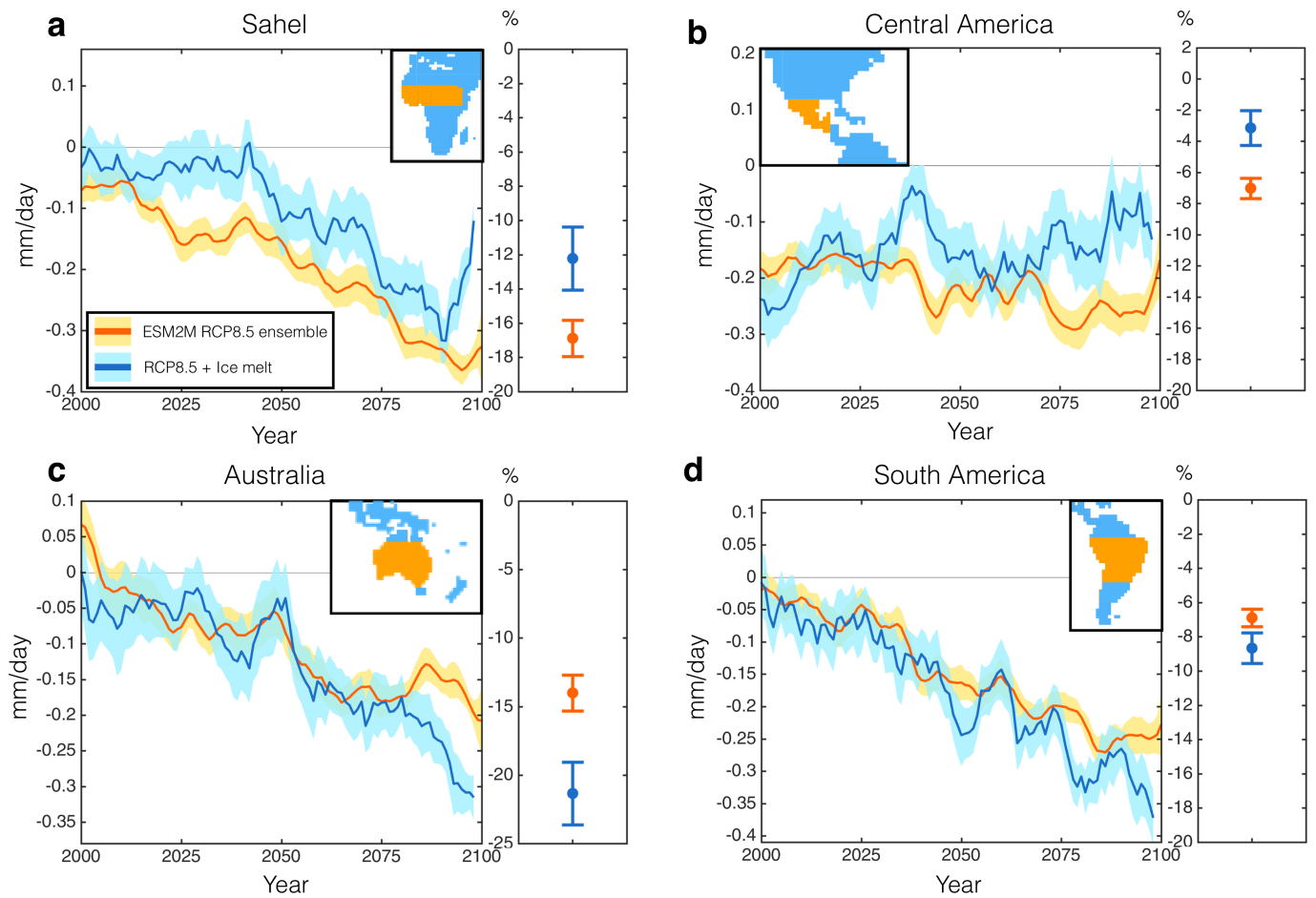
The GFDL ESM2M model code is publicly available from <https://github.com/mom-ocean>. The results from the standard- and meltwater-ensemble simulations are available from the corresponding author. The prescribed RCP8.5 freshwater flux used here is available from ref. ⁴. Antarctic sea-ice extent from satellite measurements is available from the NSIDC at <https://nsidc.org/data>. The Southern Ocean state-estimate data used for model evaluation is available from http://sosse.ucsd.edu/bosse_solution_iter105.html. The topographical data used in Figs. 1, 2, 4 and Extended Data Figs. 1, 9, 10 are available in MATLAB and provided by NOAA⁷³.

47. Gent, P. R. et al. The community climate system model version 4. *J. Clim.* **24**, 4973–4991 (2011).
48. Dunne, J. P. et al. GFDL’s ESM2 global coupled climate-carbon earth system models. Part I: physical formulation and baseline simulation characteristics. *J. Clim.* **25**, 6646–6665 (2012).

49. Dunne, J. P. et al. GFDL's ESM2 global coupled climate-carbon earth system models. Part II: Carbon system formulation and baseline simulation characteristics. *J. Clim.* **26**, 2247–2267 (2013).
50. Griffies, S. The Gent-McWilliams skew flux. *J. Phys. Oceanogr.* **28**, 831–841 (1998).
51. Stocker, T. et al. in *Climate Change 2013: The Physical Science Basis. Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Stocker, T. F. et al.) 33–115 (Cambridge Univ. Press, Cambridge, 2013).
52. Sallée, J. B. et al. Assessment of Southern Ocean water mass circulation and characteristics in CMIP5 models: historical bias and forcing response. *J. Geophys. Res. Oceans* **118**, 1830–1844 (2013).
53. Shu, Q., Song, Z. & Qiao, F. Assessment of sea ice simulations in the CMIP5 models. *Cryosphere* **9**, 399–409 (2015).
54. Reintges, A., Martin, T., Latif, M. & Park, W. Physical controls of Southern Ocean deep-convection variability in CMIP5 models and the Kiel climate model. *Geophys. Res. Lett.* **44**, 6951–6958 (2017).
55. Gordon, A. Deep Antarctic convection west of Maud rise. *J. Phys. Oceanogr.* **8**, 600–612 (1978).
56. de Lavergne, C., Palter, J. B., Galbraith, E. D., Bernardello, R. & Marinov, I. Cessation of deep convection in the open Southern Ocean under anthropogenic climate change. *Nat. Clim. Chang.* **4**, 278–282 (2014).
57. Pellichero, V., Sallée, J.-B., Schmidtke, S., Roquet, F. & Charrassin, J.-B. The ocean mixed layer under Southern Ocean sea-ice: seasonal cycle and forcing. *J. Geophys. Res. Oceans* **122**, 1608–1633 (2017).
58. Swart, N. C. & Fyfe, J. C. Observed and simulated changes in the southern hemisphere surface westerly wind-stress. *Geophys. Res. Lett.* **39**, L16711 (2012).
59. Downes, S. M. & Hogg, A. M. Southern Ocean circulation and eddy compensation in CMIP5 models. *J. Clim.* **26**, 7198–7220 (2013).
60. Frölicher, T. L. et al. Dominance of the Southern Ocean in anthropogenic carbon and heat uptake in CMIP5 models. *J. Clim.* **28**, 862–886 (2015).
61. Verdy, A. & Mazloff, M. R. A data assimilating model for estimating Southern Ocean biogeochemistry. *J. Geophys. Res. Oceans* **122**, 6968–6988 (2017).
62. Rodgers, K. B., Lin, J. & Froelicher, T. L. Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an Earth system model. *Biogeosciences* **12**, 3301–3320 (2015).
63. Wang, Z. et al. An atmospheric origin of the multi-decadal bipolar seesaw. *Sci. Rep.* **5**, 8909 (2015).
64. Meehl, G. A., Arblaster, J. M., Bitz, C. M., Chung, C. T. Y. & Teng, H. Antarctic sea-ice expansion between 2000 and 2014 driven by tropical Pacific decadal climate variability. *Nat. Geosci.* **9**, 590–595 (2016).
65. Depoorter, M. A. et al. Calving fluxes and basal melt rates of Antarctic ice shelves. *Nature* **502**, 89–92 (2013).
66. Dupont, T. & Alley, R. Assessment of the importance of ice-shelf buttressing to ice-sheet flow. *Geophys. Res. Lett.* **32**, L04503 (2005).
67. Lazeroms, W. M. J., Jenkins, A., Gudmundsson, G. H. & van de Wal, R. S. W. Modelling present-day basal melt rates for Antarctic ice shelves using a parametrization of buoyant meltwater plumes. *Cryosphere* **12**, 49–70 (2018).
68. MacAyeal, D. R. in *Oceanology of the Antarctic Continental Shelf* (ed. Jacobs, S.) 133–143 (American Geophysical Union, Washington, 1985).
69. Holland, P. R., Jenkins, A. & Holland, D. M. The response of ice shelf basal melting to variations in ocean temperature. *J. Clim.* **21**, 2558–2572 (2008).
70. Little, C. M., Gnanadesikan, A. & Oppenheimer, M. How ice shelf morphology controls basal melting. *J. Geophys. Res. Oceans* **114**, C12007 (2009).
71. Goldberg, D. N. et al. Investigation of land ice-ocean interaction with a fully coupled ice-ocean model: 2. Sensitivity to external forcings. *J. Geophys. Res. Earth Surf.* **117**, F02038 (2012).
72. Griffies, S. M. *Elements of the modular ocean model (MOM)*. Report No. 7, https://github.com/mom-ocean/mom-ocean.github.io/blob/master/assets/pdfs/MOM5_elements.pdf (NOAA GFDL Ocean Group, 2012).

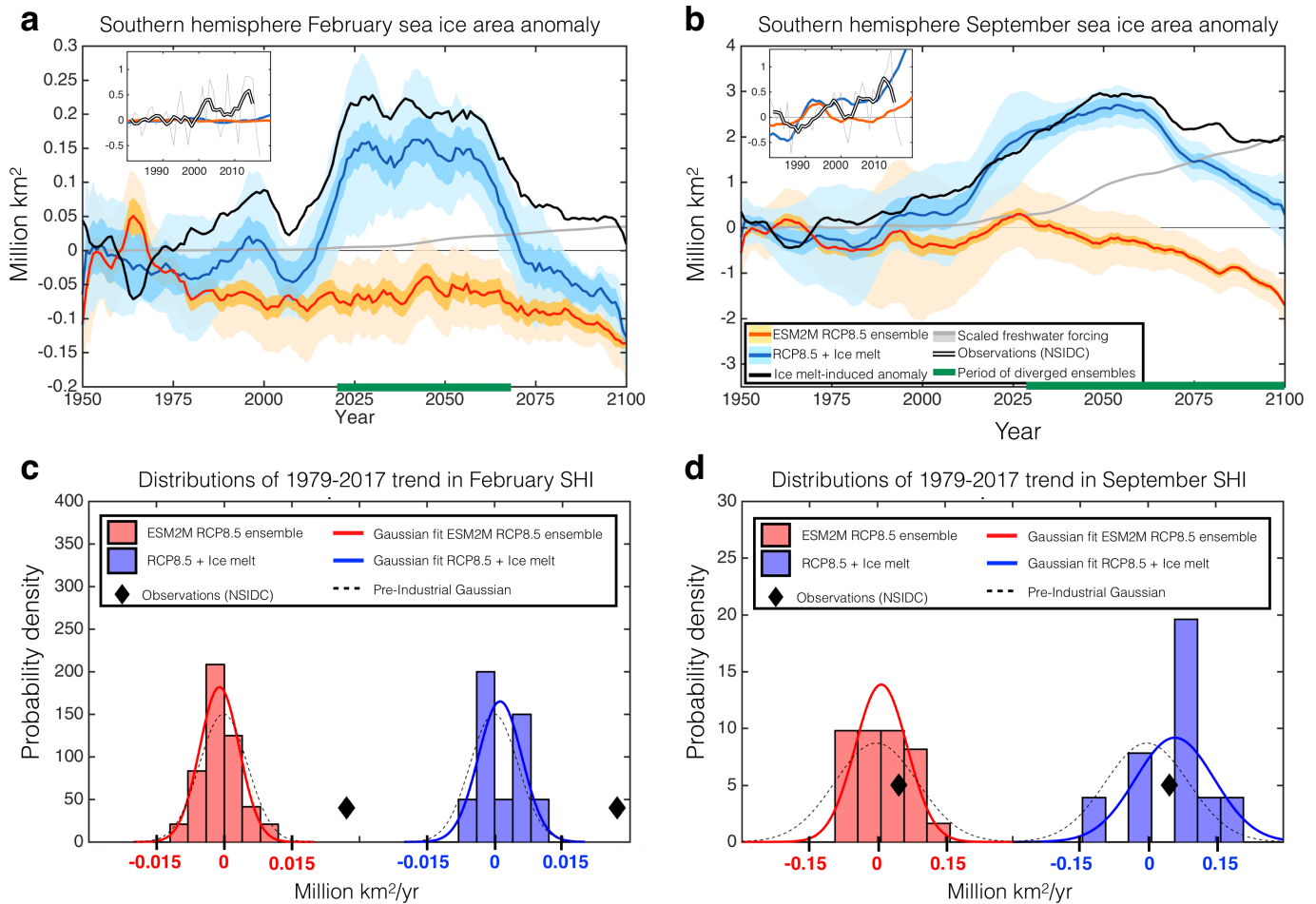


Extended Data Fig. 1 | Applied RCP8.5 freshwater flux. The orange line shows the digitized data applied to ESM2M and the blue line shows the original data⁴ (1 Sv = $10^6 \text{ m}^3 \text{ s}^{-1}$).



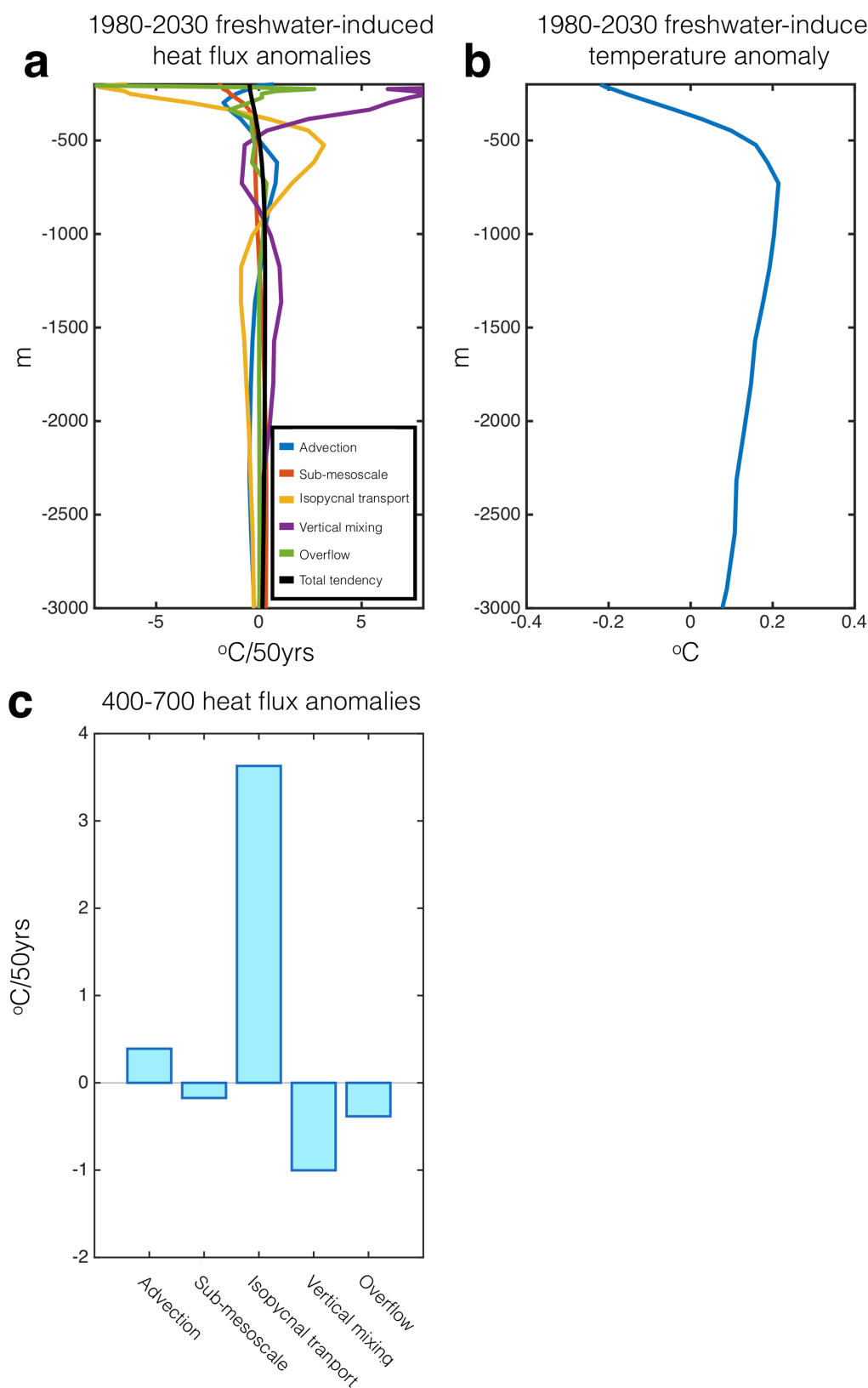
Extended Data Fig. 2 | Regional-area mean precipitation–evaporation anomalies. a–d, Anomalies are shown as a function of time for the Sahel (a), Central America (b), Australia (c) and South America (d). Orange shows the standard RCP8.5 30-member ESM2M ensemble ('standard ensemble') and blue shows the 10-member RCP8.5 with added time-varying freshwater melt around Antarctica ('meltwater ensemble'). The solid lines show the ensemble means and the shading shows the 95%

uncertainty in the mean. The data points on the right of each panel show the 2080–2100 mean anomalies, expressed as a percentage relative to the pre-industrial mean state (note the different vertical axes), with the error bars showing the 95% uncertainty in the means. Here, the anomalies are calculated with respect to the pre-industrial control simulation. The maps in the insets indicate the area over which the respective anomalies (colour-coded) are calculated. All time series are smoothed with a 5-year filter.



Extended Data Fig. 3 | Seasonal sea-ice anomalies. **a, b**, Time series of the February (**a**) and September (**b**) SHI anomalies relative to the 1950–1970 mean. Orange shows the standard ensemble and blue shows the meltwater ensemble. Solid lines show ensemble means, the dark shading shows the uncertainty in the mean and the light shading shows the full ensemble spread of 20-year SHI means. The solid black line shows the difference between the orange and blue lines, and the applied meltwater flux is shown in grey (scaled to the final 5-year mean of the meltwater-induced SHI anomaly). The green bar indicates the period when the full ensembles have diverged. The insets show the period 1980–2020, with the

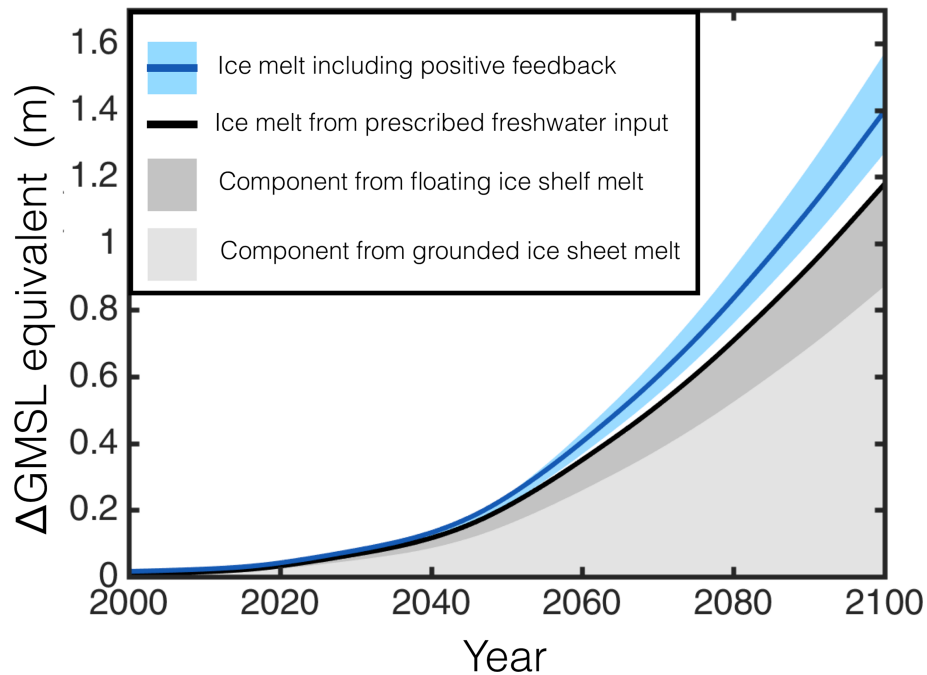
double black line showing the observed monthly mean sea-ice area from the NSIDC, relative to the 1980–2000 mean. The thin grey line shows the unsmoothed observations. **c, d**, Distribution of linear trends in SHI over the period 1979–2017, calculated for each ensemble member, for February (**c**) and September (**d**) means. The red bars show the standard ensemble and blue bars show the meltwater ensemble, with different x axes. The solid lines show Gaussian fits to the distributions and the dashed black line shows the pre-industrial distribution. The observations are shown as black diamonds.



Extended Data Fig. 4 | Heat-budget analysis. a, b, 1980–2030 freshwater-induced heat-flux anomalies (**a**) and temperature anomaly (**b**), averaged along the Antarctic coast, due to a 0.1-Sv freshwater perturbation, as a function of depth. **c**, 1980–2030-mean freshwater-induced heat-flux

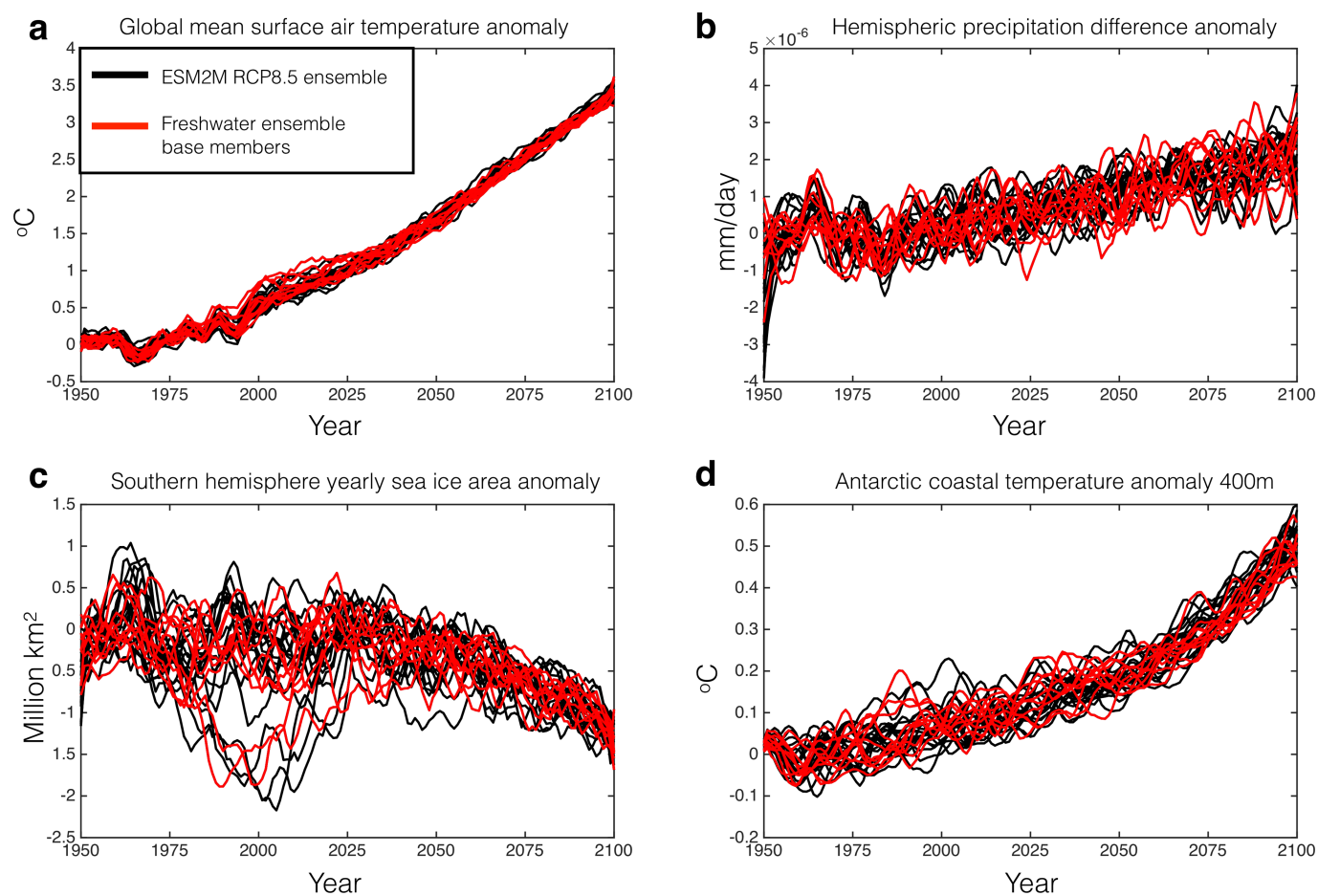
anomalies, averaged over 400–700-m depth along the Antarctic coast. All anomalies shown here are calculated relative to the mean of standard ensemble.

Freshwater input



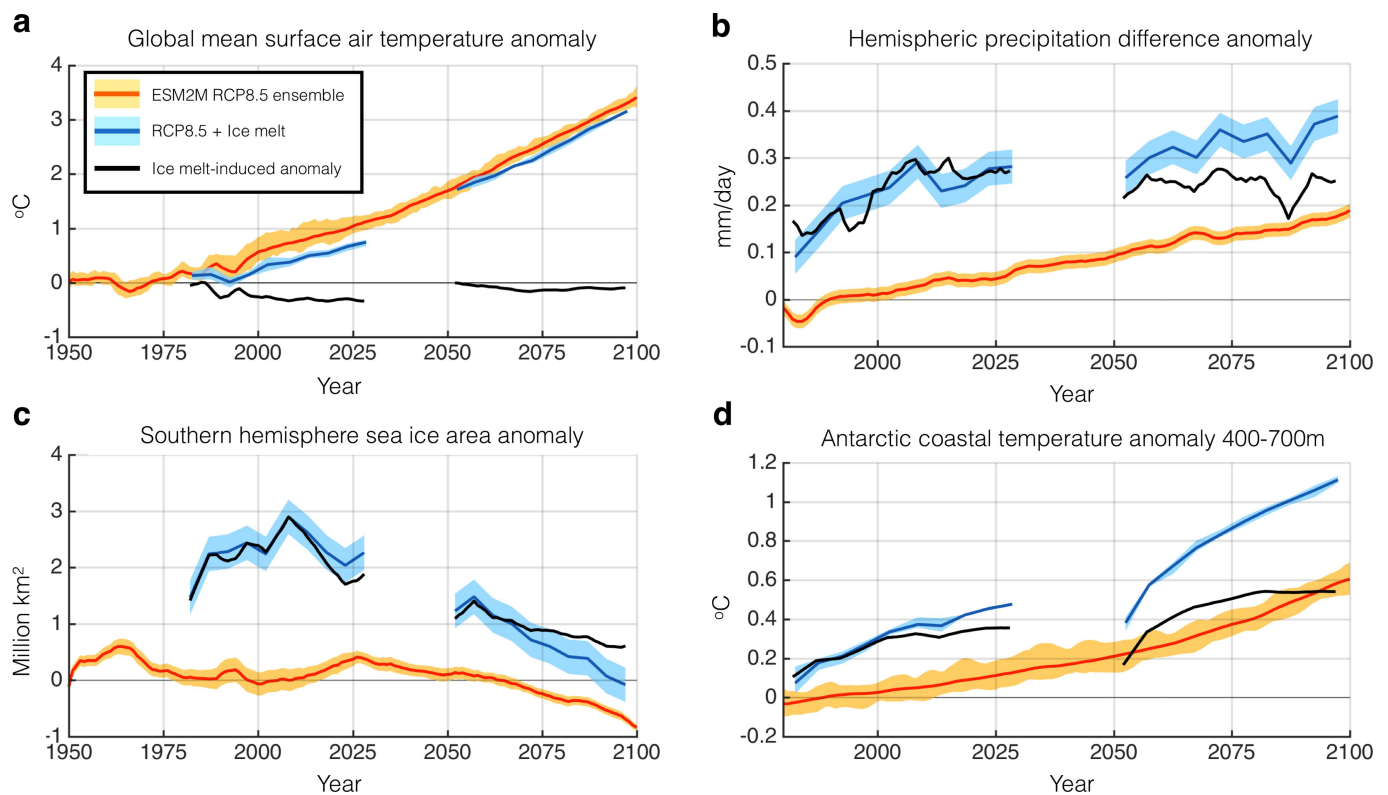
Extended Data Fig. 5 | Ice-melt feedback. The ice-melt freshwater input, in equivalent global-mean sea level (ΔGMSL), due to the RCP8.5 prescribed meltwater is shown in black. The dark and light grey shading show the components of the prescribed flux from ice-shelf and ice-sheet melt, respectively. Only the ice-sheet melt contributes to sea level. The

blue line shows the total freshwater flux, including the prescribed flux and the estimated feedback associated with ice-shelf melt from the freshwater-induced ocean warming. The blue shading shows the 95% uncertainty range.



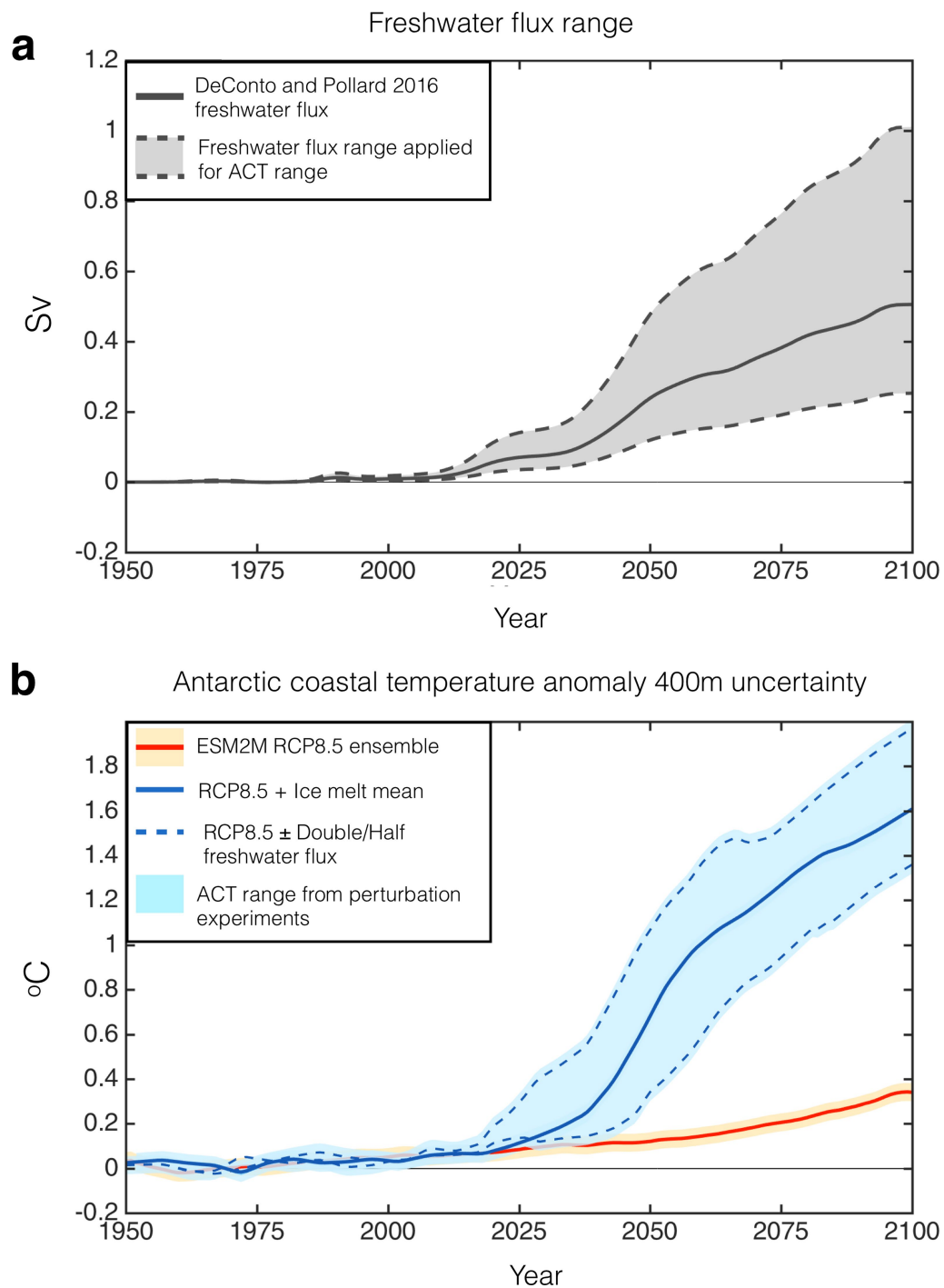
Extended Data Fig. 6 | Selection of ensemble members for meltwater experiments. a–d, Time series of the global-mean SAT (**a**), PRE (**b**), annual-mean SHI (**c**) and ACT (**d**) anomalies in the standard ensemble

relative to the pre-industrial control. The black lines show all 30 ensemble members and the red lines show those used for the freshwater-forced simulations.



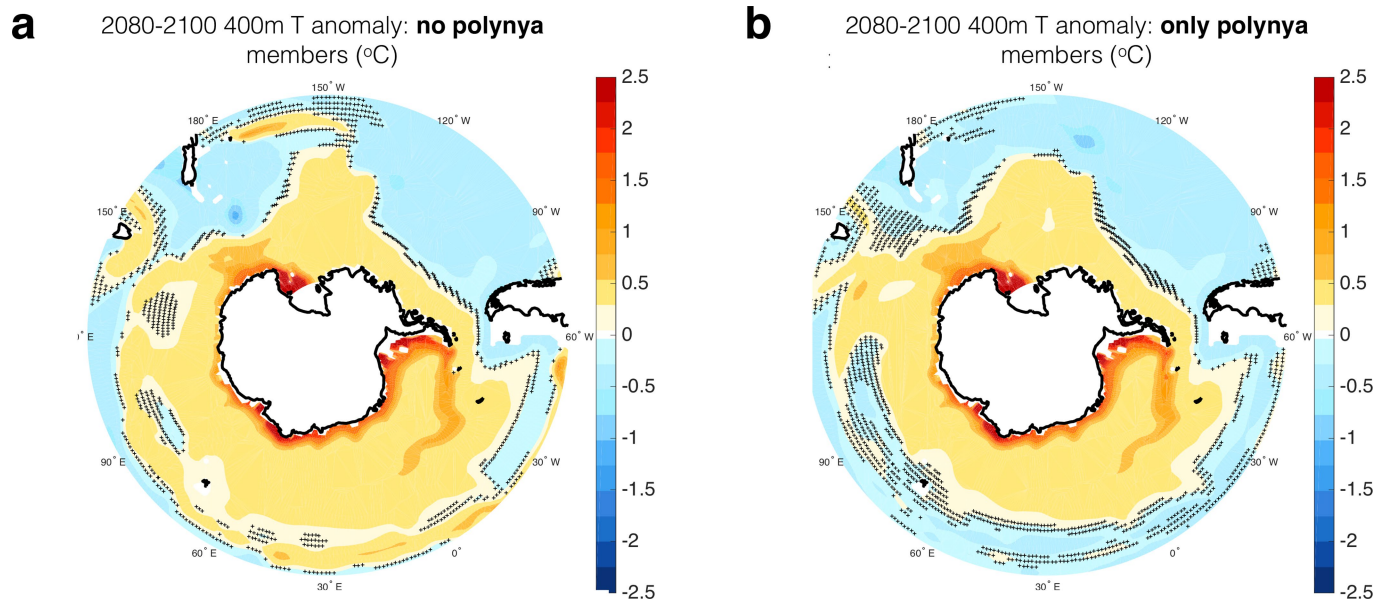
Extended Data Fig. 7 | Sensitivity to base state. **a–d**, Time series of the global-mean SAT (**a**), PRE (**b**), SHI (**c**) and ACT (**d**) anomalies relative to the pre-industrial control. Orange shows the yearly standard ensemble and blue shows the 5-year means of the meltwater ensemble. The meltwater ensemble in these experiments is hosed with 0.1 Sv for 50 years in two

separate periods, 1980–2030 and 2050–2100, initialized from the standard ensemble. Solid lines show ensemble means and the dark shading shows the 90% uncertainty in the mean. The solid black lines show the difference between the meltwater and standard ensembles.

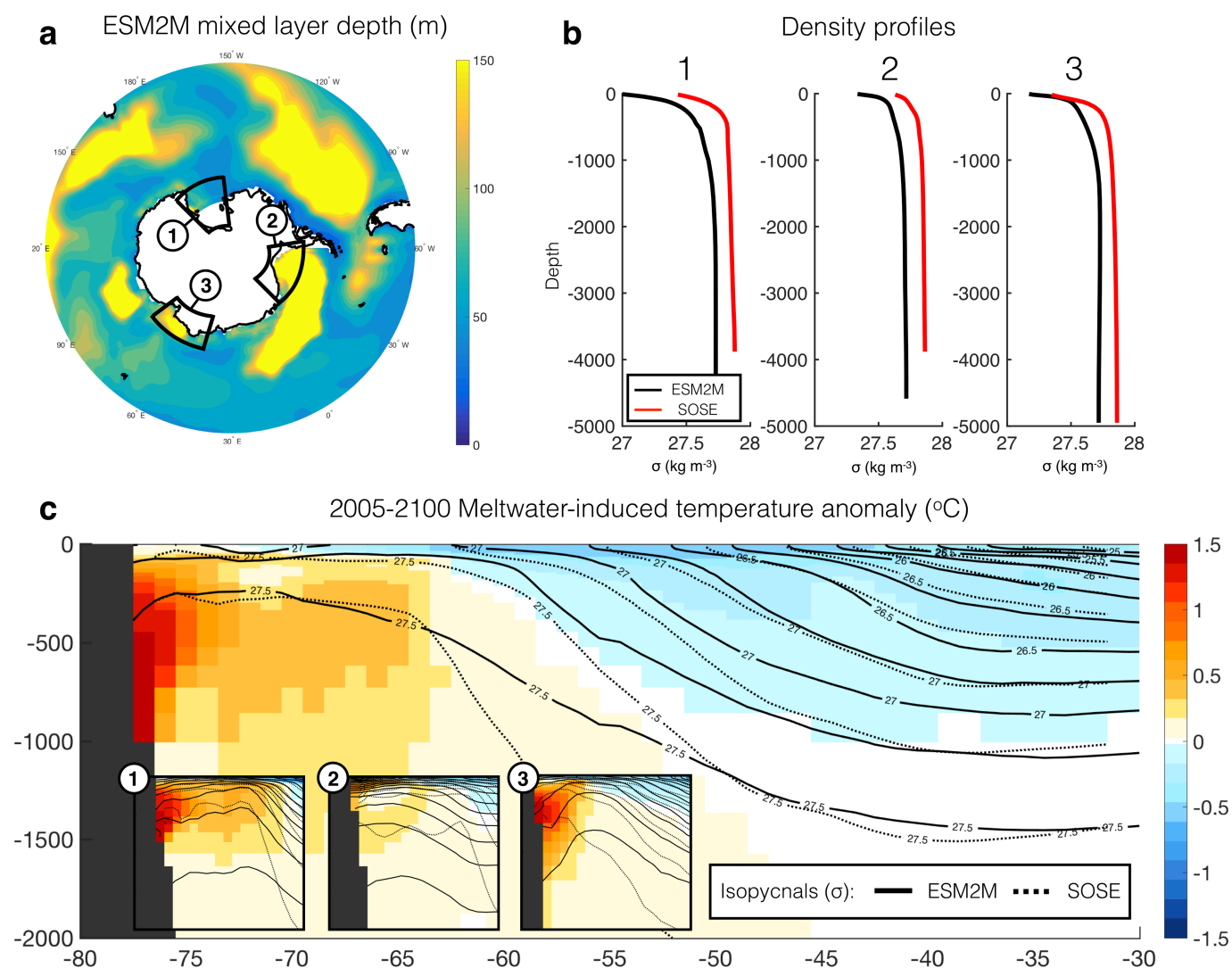


Extended Data Fig. 8 | Range of 400-m ocean warming. a, Range of freshwater flux. The solid grey line shows the projected flux⁴; the dashed lines and shading show the same flux, but multiplied by factor of 0.5 and 2. **b,** ACT anomalies in the standard (orange) and meltwater (blue) ensembles. The solid blue line shows the response to the projected

flux⁴; the dashed blue lines show the temperature range covered by the experiments with half and double the projected flux, which have three ensemble members each. The shading shows the full ensemble spread of 20-year means.



Anomalies are calculated pair-wise, relative to the standard-ensemble members. Hatching indicates where the anomalies are not significant at the 95% level.



Extended Data Fig. 10 | ESM2M–SOSE comparison. **a**, ESM2M pre-industrial annual-mean depth of the mixed layer. **b**, Area-mean density profiles for ESM2M (black) and SOSE (red) for each of the numbered boxes in **a**. **c**, ESM2M 2005–2100-mean meltwater-induced temperature anomaly (zonal mean; colour scale), and zonal-mean ESM2M (solid) and

SOSE (dashed) isopycnal surfaces, as functions of depth and latitude. The insets illustrate these quantities for the numbered regions from **a**, showing the upper 2,000 m of the ocean, between 80° S and 60° S (regions (1) and (2)) or 70° S and 50° S (region (3)).

Abiotic synthesis of amino acids in the recesses of the oceanic lithosphere

Bénédicte Ménéz^{1*}, Céline Pisapia^{1,2}, Muriel Andreani^{3*}, Frédéric Jamme², Quentin P. Vanbellingen⁴, Alain Brunelle⁴, Laurent Richard⁵, Paul Dumas² & Matthieu Réfrégiers²

Abiotic hydrocarbons and carboxylic acids are known to be formed on Earth, notably during the hydrothermal alteration of mantle rocks. Although the abiotic formation of amino acids has been predicted both from experimental studies and thermodynamic calculations, its occurrence has not been demonstrated in terrestrial settings. Here, using a multimodal approach that combines high-resolution imaging techniques, we obtain evidence for the occurrence of aromatic amino acids formed abiotically and subsequently preserved at depth beneath the Atlantis Massif (Mid-Atlantic Ridge). These aromatic amino acids may have been formed through Friedel–Crafts reactions catalysed by an iron-rich saponite clay during a late alteration stage of the massif serpentinites. Demonstrating the potential of fluid–rock interactions in the oceanic lithosphere to generate amino acids abiotically gives credence to the hydrothermal theory for the origin of life, and may shed light on ancient metabolisms and the functioning of the present-day deep biosphere.

Abiotic synthesis of organic compounds by the reduction of inorganic carbon species is thermodynamically favoured by the production of molecular hydrogen (H₂), which accompanies serpentinization reactions¹. In these hydration reactions, the production of H₂ results from the reduction of water coupled to the oxidation of ferrous iron in olivine and pyroxene, the major rock-forming minerals of the upper mantle^{1,2}. Therefore, hydrothermal areas where active serpentinization occurs are increasingly regarded as possible settings for the appearance of the first building blocks of life and the emergence of primordial metabolisms^{3,4}. In that perspective, the discovery of the Lost City hydrothermal field hosted on the Atlantis Massif near the Mid-Atlantic Ridge⁵ profoundly changed our vision of Earth habitability, as it provides a modern example of H₂-rich alkaline fluids generated at moderate temperatures (50–150 °C)—that is, the most favourable conditions for the appearance of life⁶.

During the past two decades, experimental studies and thermodynamic calculations outlined the potential of serpentinization reactions to promote the abiotic formation of a diversity of organic compounds, including some of direct interest to prebiotic chemistry such as the protein-forming amino acids (Supplementary Table 1). However, reaction pathways for the abiotic formation of amino acids under hydrothermal conditions are still not well constrained, although several mechanisms have been proposed. Among the most commonly invoked reactions, the Strecker synthesis produces amino acids (mostly aliphatic) from the reaction of aldehydes or ketones with ammonia and cyanide. The overall reaction is thermodynamically favoured in hydrothermal environments and possibly catalysed by metals and minerals (Supplementary Table 1).

Geochemical and isotopic evidence demonstrates the abiotic synthesis of only a restricted number of low-molecular-mass organic compounds in hydrothermal systems associated with serpentinization on Earth. These compounds include mainly methane, short-chain alkanes and formate⁷. Although the presence of amino acids has been reported in hydrothermal vent fluids, these amino acids probably derived from ecosystems hosted in the shallow oceanic crust⁸. Primordial amino acids and other nitrogen-bearing organic

compounds have generally been considered as extra-terrestrial in origin, and exogenously delivered to Earth by comets and asteroids⁹. These organic nitrogen compounds would be inherited from the aqueous alteration of the chondrite parent bodies or the asteroidal meteorites themselves, in a process that resembles serpentinization as it occurs on Earth^{10,11}. Although it was previously acknowledged⁸ that some of the amino acids detected at the Lost City hydrothermal field could have been synthesized abiotically, the possibility that the serpentinizing oceanic lithosphere could represent an efficient factory for nitrogen-bearing organic compounds has been poorly assessed until now. The current approaches are mainly based on the analysis of fluids discharged at hydrothermal vents where, if abiotic synthesis would occur, products are likely to be too diluted to be distinguished from background biological contamination⁸.

Serpentinite-hosted amino acids

Our study focuses on a deeply serpentinized harzburgite that was recovered by drilling in the Atlantis Massif at a depth of 173.15 m below sea floor during the Integrated Ocean Drilling Program (IODP) Expedition 304 at Hole U1309D¹². The Atlantis Massif is a tectonically exhumed dome associated with an oceanic core complex located at the intersection between the Mid-Atlantic Ridge and the Atlantis transform fault (30° 8' N–42° 8' W). The sample was selected based on its high content in organic carbon (that is, 232 p.p.m.)¹³. The high-temperature (300–350 °C)¹² hydrated paragenesis resembles that usually found in serpentinized peridotites, with olivine being replaced by serpentine and magnetite exhibiting a characteristic mesh texture (Extended Data Fig. 1). Yellow to brownish phases are frequently found in the core of the mesh serpentine (Fig. 1a, Extended Data Fig. 1). These phases correspond to Fe-rich serpentine and Fe-rich saponite enriched in organic carbon^{14,15} (Extended Data Figs. 2–4f). They formed, respectively, during a second and third stage of hydrothermal alteration that occurred at lower temperature (less than 200 °C for Fe-rich serpentine¹⁶ and less than 100–150 °C for saponite¹⁷) at the expense of the first generation of serpentine and of the olivine kernels (Extended Data Fig. 1).

¹Institut de Physique du Globe de Paris, Sorbonne Paris Cité, Université Paris Diderot, CNRS, Paris, France. ²Synchrotron SOLEIL, Gif-sur-Yvette, France. ³Laboratoire de Géologie de Lyon: Terre, Planètes, Environnement, UMR5276, ENS-Université Lyon I, Villeurbanne, France. ⁴Institut de Chimie des Substances Naturelles, CNRS UPR2301, Université Paris-Sud, Université Paris-Saclay, Gif-sur-Yvette, France. ⁵Nazarbayev University, School of Mining & Geosciences, Astana, Kazakhstan. *e-mail: menez@ipgg.fr; muriel.andreani@univ-lyon1.fr

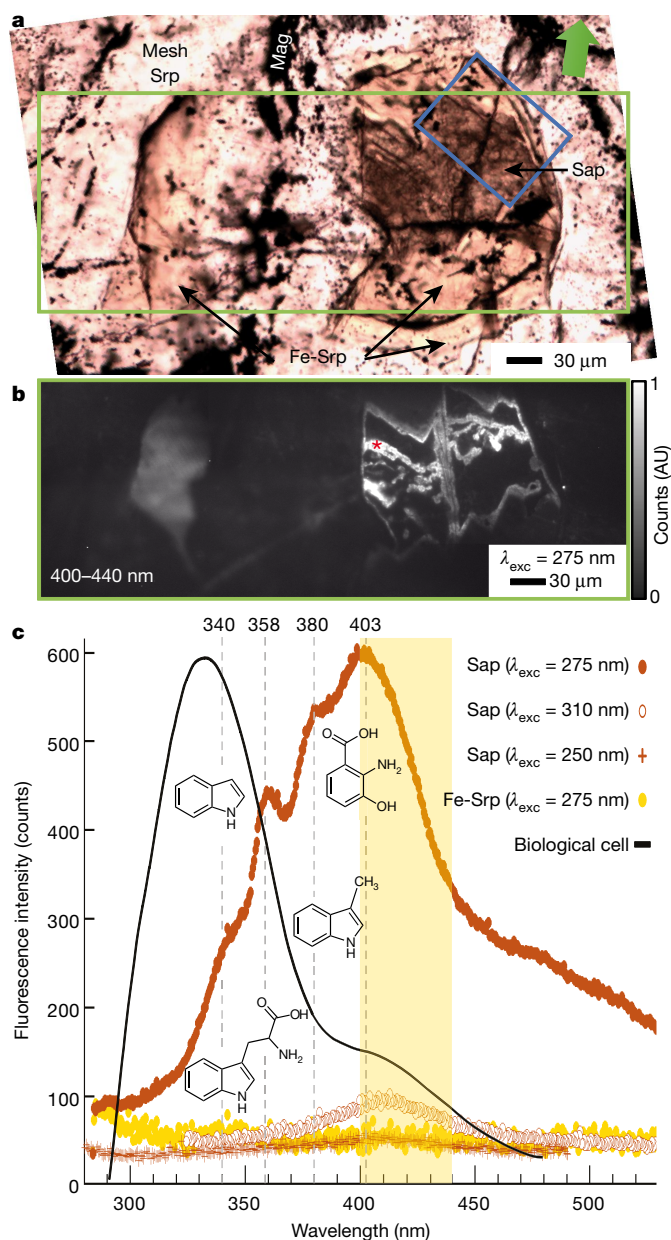


Fig. 1 | Endogenous UV-autofluorescence locally revealed by S-DUV imaging of a highly altered mantle-rock recovered at 173.15 m below sea floor from IODP Expedition 304 Hole U1309D. UV-fluorescence emerged from heteroatomic aromatic compounds shown to be spatially restricted to an Fe-rich clay in which they are heterogeneously distributed. **a**, Optical image showing yellow to brownish phases identified as Fe-rich serpentine (Fe-Srp) and Fe-rich saponite (Sap), hosted in a serpentinized harzburgite with olivine being replaced by magnetite (Mag) and serpentine exhibiting a characteristic mesh texture (mesh Srp); the green arrow indicates sample orientation. **b**, Full field S-DUV image of the area depicted by the light-green rectangle in **a**, collected between 400 and 440 nm using excitation (λ_{exc}) at 275 nm. AU, arbitrary units. **c**, Fluorescence emission spectra collected with excitation wavelengths of 250, 275 and 310 nm at the location shown by an asterisk in **b**. The spectra collected at 250 and 310 nm do not show a notable UV-autofluorescence signal, whereas the spectrum at 275 nm displays fluorescence characteristic of indole at 340 ± 6 nm, tryptophan at 358 ± 3 nm, skatole at 380 ± 3 nm, and hydroxyanthranilic acid at 403 ± 3 nm¹⁸ (mean \pm s.d. of three independent fits performed on three areas). Also shown are a fluorescence emission spectrum collected at 275 nm in the Fe-rich serpentine and the typical emission spectrum of a biological cell showing maximum fluorescence emission, mainly arising from protein-forming tryptophan, shifted to 335 nm²¹. The orange area in **c** represents the fluorescence detection range used in **b**. The blue box in **a** indicates the location of complementary S-FTIR measurements (Extended Data Fig. 3).

Synchrotron-coupled deep-ultraviolet (S-DUV) microspectroscopy with excitation in the range of 250–310 nm revealed intense UV-autofluorescence (Fig. 1, Extended Data Fig. 5) where Fe-rich saponite was present (Extended Data Figs. 2–4). Full-field imaging of the fluorescence collected between 400 and 440 nm after excitation at 275 nm showed a heterogeneous spatial distribution of the strongest fluorescence intensities that formed a tortuous network within Fe-rich saponite (Fig. 1b). This is consistent with scanning electron microscopy (SEM) observations that highlight variable content in the organic carbon trapped in saponite (Extended Data Fig. 2a, b). S-DUV autofluorescence was weaker in Fe-rich serpentine than in Fe-rich saponite and nearly absent in adjacent mesh serpentine (Fig. 1). The fluorescence signal collected between 300 and 550 nm after excitation at 275 nm was characterized by four broad and overlapping bands centred at 340 ± 6 , 358 ± 3 , 380 ± 3 and 403 ± 3 nm (mean \pm s.d.) (Fig. 1, Extended Data Fig. 5). These spectral features are indicative of the presence of tryptophan to which the band at 358 nm can be assigned, and of indole, skatole and hydroxyanthranilic acid, the fluorescence spectra of which were reported previously¹⁸. These three latter compounds may correspond to products of either natural¹⁸ or UV-induced degradation¹⁹ of tryptophan, although indole can also be an intermediate in the abiotic synthesis of tryptophan²⁰. The four organic compounds were always spatially associated at comparable relative intensities (Extended Data Fig. 5). In contrast to the fluorescence emission observed after excitation at 275 nm, which is close to the maximum absorption wavelength of tryptophan²¹, excitations at 250 and 310 nm did not lead to any endogenous fluorescence emission (Fig. 1c). Whereas protein-forming tryptophan in biological cells fluoresces at 335 nm after excitation at 275 nm and does not produce notable signals at higher wavelengths²¹ (Fig. 1c), the UV-fluorescence emission value obtained here for tryptophan agrees with the maximum fluorescence emission that arises when this amino acid is free (that is, 360 nm)²¹, with the spectral shift being due to environment-related effects.

In agreement with S-DUV microspectroscopy, time-of-flight secondary ion mass spectrometry (TOF-SIMS) imaging recorded in Fe-rich saponite revealed a systematic presence of fragment ions that were characteristic of tryptophan²² (for example, $\text{C}_9\text{H}_8\text{N}^+$; Fig. 2b, c and Supplementary Table 2). In addition to saponite/Fe-rich serpentine assemblages, tryptophan was also detected inside saponite close to olivine kernels, although it was more spatially restricted and at lower fluorescence intensities (Extended Data Figs. 6, 7). In all the areas where saponite and fragment ions characteristic of tryptophan were detected, TOF-SIMS analysis did not provide any evidence for the presence of biomarkers, such as hopanoids, cholestane, pristane, squalane, lycopane or β -carotane^{23–25}, which are constituents of marine-dissolved organic carbon¹³ or of deep microbial communities (Fig. 2d, Extended Data Fig. 8).

The presence of N-bearing organic compounds in the Fe-rich saponite was confirmed by synchrotron-Fourier-transform-infrared microspectroscopy (S-FTIR), with vibrational modes attributable to pyrrole and aromatic rings (at $1,380\text{ cm}^{-1}$ and $1,460\text{--}1,465\text{ cm}^{-1}$), α -amines (between $1,460\text{--}1,465\text{ cm}^{-1}$ and $1,550\text{--}1,650\text{ cm}^{-1}$) and carboxyl functional groups (at $1,412\text{ cm}^{-1}$ and $1,728\text{ cm}^{-1}$) (Extended Data Fig. 3b, c, Supplementary Table 3). The absorption band distribution resembles that observed for tryptophan (<https://webbook.nist.gov/cgi/cbook.cgi?ID=C73223&Mask=80>), although with higher proportions of primary amines and heterocycles and higher aliphaticity. Consequently, the broad absorption band between $1,550$ and $1,650\text{ cm}^{-1}$ probably represents several overlapping bands from different N-bearing heterocycles, as also observed by S-DUV microspectroscopy. No absorption bands were detected between $1,627$ and $1,670\text{ cm}^{-1}$, in which the characteristic amide I absorption band of protein secondary structures is observed²⁶.

An abiotic origin for the amino acids

The unique organic signatures derived from S-DUV, S-FTIR and TOF-SIMS measurements, all preferentially associated with Fe-rich saponite,

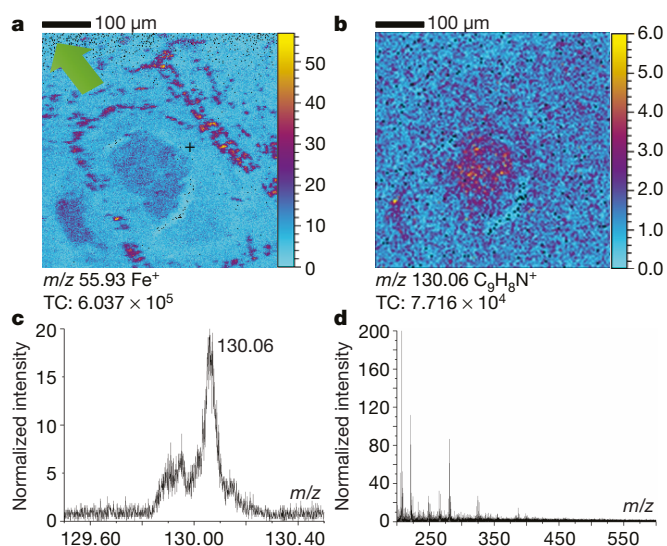


Fig. 2 | The presence of tryptophan in the Fe-rich saponite was confirmed by TOF-SIMS imaging with the co-localized collection of its characteristic fragment ions. a, TOF-SIMS ion image of Fe^+ showing the location of analysis with respect to the mesh for the area displayed in Fig. 1a. The green arrow indicates sample orientation. TC, total count. **b**, **c**, Molecular ion image (binning of four) of the $\text{C}_9\text{H}_8\text{N}^+$ fragment ion at a m/z of 130.06 (**b**), and the corresponding ion peak (**c**), which is characteristic of tryptophan²². It reveals cluster-like accumulations within the Fe-rich saponite. **d**, TOF-SIMS spectrum (m/z of 200–600) reconstructed from the region displaying the highest count rates in **b**. It provides evidence for the absence of common biomarkers^{23–25}. Detailed spectra are provided in Extended Data Fig. 8a–n. See Supplementary Table 2 for full list of TOF-SIMS fragment ions related to tryptophan that were detected in the region displaying the highest count rates in **b**.

are clearly different from those obtained under comparable analytical conditions for microbial cells or biofilm-forming extracellular polymeric substances (Figs. 1c, 2, Extended Data Figs. 3b, 8). The presence of microorganisms or their remnants would have resulted in a spatially variable complex mixture of biopolymers, all carrying diverse functional groups²⁷. By contrast, the organic material detected here corresponds to low-molecular-mass compounds with TOF-SIMS m/z of less than 350 (Fig. 2d, Extended Data Fig. 8). In addition, S-DUV, S-FTIR and TOF-SIMS analyses constantly display spectral signatures that vary little from one micrometre to the other, all in favour of an abiotic origin for the tryptophan. In agreement with this hypothesis, as observed by transmission electron microscopy (TEM) analysis, the Fe-rich saponite nanoporous network that hosts the tryptophan signal is too small to host or have hosted prokaryotic cells of a few micrometres in length (Fig. 3a, b, Extended Data Figs. 4, 6, 7). The lack of a characteristic amino acid signal in microfractures related to the Fe-rich saponite is in support of tryptophan endogeneity.

At the nanoscale, the C-enriched Fe-rich saponite displayed highly variable texture and porosity (Fig. 3a, b, Extended Data Fig. 6d). At the interface between olivine crystals and mesh serpentine or Fe-rich serpentine—that is, where organic compounds displayed weak UV-fluorescence and TOF-SIMS signals (Extended Data Fig. 6)—Fe-rich saponite lamellae are mainly subparallel although some sheet distortions are visible (Extended Data Figs. 6d, 7d). Where tryptophan was found at higher concentrations in assemblages of Fe-rich saponite and Fe-rich serpentine (Figs. 1, 2, Extended Data Figs. 2–5), TEM analysis showed the presence of packed saponite sheets in face–face mode, edge–edge mode or edge–face mode, forming a nanoporous house-of-cards structure (Fig. 3a, b). Such a state of structural disorder is supported by the lack of absorption at $3,415\text{ cm}^{-1}$ in the Fe-rich saponite S-FTIR spectra, a band that theoretically corresponds to interlamellar water²⁸ (Extended Data Fig. 3c). Together, these observations suggest the formation of a reactive organoclay²⁹ during the late stages of aqueous alteration of the Atlantis Massif serpentine.

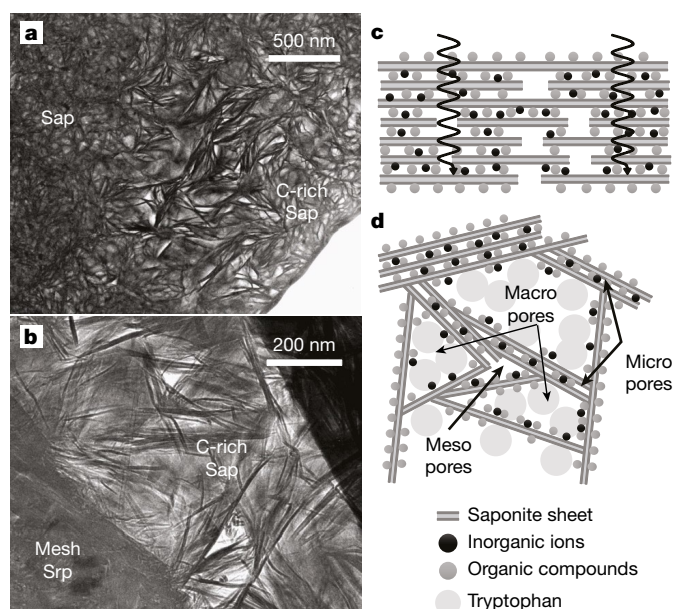


Fig. 3 | Fe-rich saponite as a catalysing pillared-clay for the abiotic synthesis of tryptophan. a, b, TEM images collected on the focused ion beam (FIB) foil milled in the area shown in Figs. 1, 2 and Extended Data Figs. 2–5. Saponite shows the presence of aggregates formed by the packing of sheets opening nanopores, and suggests the occurrence of pillaring processes and a high reactivity of the clay. **c**, Clay minerals possess negatively charged silicate layers with cations in the interlayer space to balance the charge. They allow sorption and the exchanges of ions or organic moieties to occur during water–clay interactions, hence propping apart clay layers at increased interlayer distances, that is, packing in face–face mode. **d**, Packing in face–face, edge–edge or edge–face mode results in a house-of-cards structure²⁹ that is recognizable in **a** and **b**. The nanoporous structure represents microreactors for the formation of tryptophan at the acid sites of the Fe-rich saponite.

Abiotic Friedel–Crafts–type synthesis

Saponite, a tetrahedrally charged trioctahedral smectite, is a well-known acid catalyst with high adsorption, swelling and cation-exchange capacities²⁸. These properties are widely used in the industry for organic synthesis or for the reduction of inorganic nitrogen species^{28,29}. Some varieties of smectite, including saponite, were shown to promote abiotic synthesis of (poly)aromatic hydrocarbons under hydrothermal conditions (that is, 300°C , 100 MPa)³⁰. The properties of smectite enable the formation of pillared structures out of their expandable silicate sheets due to their ability to exchange with interlayer water and insert in their structure ionic compounds and/or organic moieties that act as pillars that prop apart the clay sheets at increased interlayer distances (Fig. 3c), as observed by TEM analysis (Extended Data Fig. 6d). Further stacking and distortion of the clay lamellae creates a network of interconnected nanopores with increased surface area and reducing properties where organosynthesis can proceed and of which the products enhance interlayer expansion²⁹ (Fig. 3a, b, d).

In the general structure of saponite, the octahedral sites are usually occupied by Mg^{2+} ions, and Al^{3+} ions substitute for Si in the tetrahedral sites. In the iron-bearing variety, both Fe^{2+} and Fe^{3+} cations replace Mg^{2+} and Al^{3+} (Supplementary Tables 4 and 5), thus conserving the negative charge of the silicate sheets and hence the cation-exchange capacities while maintaining the layered structure³¹. The presence of Fe^{3+} in the tetrahedral sheets (Supplementary Table 5) also enhances surface acidity and therefore sorption and catalysis properties of saponite³¹. The existence in the silicate sheets of Brønsted and Lewis acid sites to which aromatic and heteroatomic compounds may sorb is suggested by the multiple absorption bands observed between $1,550$ and $1,650\text{ cm}^{-1}$ in the Fe-rich saponite S-FTIR spectrum^{28,32} (Extended Data Fig. 3b).

Owing to their pillaring effect, enhanced sorption capacity and high reducible iron content, Fe-smectites have been shown to be the most efficient solid catalyst for Friedel–Crafts reactions³³. Friedel–Crafts-type reactions are the method of choice in the industry for the alkylation of (hetero)arenes under the catalytic effect of Lewis or Brønsted acids, with or without co-catalysts³⁴. In the present case, Friedel–Crafts-type reactions may represent an attractive explanation for the formation of aromatic amino acids such as tryptophan. In the commonly considered Strecker synthesis⁷, aliphatic amino acids are predominantly formed (Supplementary Table 1) and additional reaction steps are required to form aromatic amino acids. By contrast, starting from aromatic hydrocarbons or heteroaromatic compounds to which substituents are added, Friedel–Crafts reactions may offer a more direct route towards the formation of tryptophan at the acid sites of Fe-rich saponite, as demonstrated experimentally through the asymmetric alkylation of indole with various catalysts²⁰. Also in favour of such a hypothesis is the considerable thermodynamic potential for the hydrothermal synthesis of (poly)aromatic hydrocarbons during serpentinization³⁵. Polyaromatic hydrocarbons may hence be available as possible reactants in the shallow oceanic lithosphere.

We therefore propose that Friedel–Crafts-type reactions may be responsible for the formation of abiotic aromatic amino acids during the hydrothermal alteration of oceanic peridotites, with this formation being catalysed by Fe-rich saponite. Possible pathways could involve the alkylation of indole with pyruvate followed by amination, with the synthesis of pyruvate under hydrothermal conditions being demonstrated experimentally³⁶. In addition, as detailed in the Supplementary Information and Extended Data Fig. 9, chemical affinity calculations using concentrations reported for the Lost City hydrothermal fluid suggest that the abiotic formation of indole, pyruvate and tryptophan from HCO_3^- , H_2 and NH_3 is thermodynamically favourable under the temperature, pH and redox conditions that prevail in the Atlantis Massif^{37,38}. The presence of NH_3 as the dominant form of nitrogen in the Atlantis Massif is supported by additional thermodynamic calculations (Extended Data Fig. 10) and the general recognition that crustal nitrogen-reduction reactions operate during hydrothermal circulation, turning oceanic nitrite and nitrate, and mantle N_2 into stable NH_3 and NH_4^+ , depending on the pH^{39,40}.

Implications

The results reported here clearly indicate that the clay-forming hydrothermal alteration of oceanic rocks has a fundamental role in the synthesis and stabilization of complex organic compounds such as aromatic amino acids. This may have far-reaching implications for the carbon and nitrogen cycles in the Earth's system, as well as for the potential for prebiotic chemistry on Earth and the deep biosphere.

Although little is known about the concentration of inorganic nitrogen species in Lost City hydrothermal fluids⁵, our discovery indicates that these fluids transport sufficient nitrogen for the abiotic synthesis of N-bearing organic compounds. In addition, our observations may extend the ranges of depths and temperatures beyond those that are generally considered as compatible for the formation and preservation of organic molecules of prebiotic interest (that is, in deep subsurface versus at hydrothermal vents). In a reduced rocky environment isolated from the open ocean and atmosphere, sorptive mechanisms can also protect the clay-trapped amino acids from hydrolysis⁴¹, as proposed for carbonaceous meteorites⁴². However, further petrological, geochemical and thermodynamic investigations are needed to better constrain the conditions under which Fe-rich saponite formed in the Atlantis Massif, with saponite precipitation occurring over a relatively large temperature range (that is, 25–200°C)⁴³.

Nonetheless, saponite constitutes the main product of the alteration of basalts and ultramafic rocks by silica-enriched fluids⁴³. The ability of Fe-rich saponite to promote and preserve precursors of biopolymers may have contributed to prebiotic synthesis when abundant (ultra) mafic rocks formed the undifferentiated primeval lithosphere covered by the large Hadean ocean⁴⁴. Although the diversity of amino acids that may have been synthesized in such a context needs to be further

explored, the scheme proposed here offers a powerful mechanism to drive the synthesis of prebiotic compounds under realistic Archaean conditions. Concentration and some forms of condensation, polymerization and further chemical evolution are possible in the chemically reducing nanopores formed by the Fe-rich saponite sheets acting as a confined microreactor (Fig. 3). Although tryptophan may not have served as a first protein building block⁴⁵, amino acids are also known to serve as biochemical precursors, deemed to catalyse the synthesis of sugars, aldehydes and nucleotide intermediates⁴⁶. If hydrothermal vent chimneys received a lot of attention for their vast network of microcompartments walled by catalytic minerals, allowing the concentration of organic synthesis products on the early Earth³, Fe-rich saponite with its tiny protective niches offers comparable attractiveness. Its ion-exchange capability drives chemical gradients and non-equilibrium conditions, and the silicate layers may play the interface part of cell membranes⁴⁷.

Finally, the possibility of abiotically formed amino acids in the recesses of the oceanic lithosphere also has important consequences for ancestral metabolisms and microbial life strategies in the present-day deep biosphere. Both are strongly linked to the nature of compounds that can be used as carbon and energy sources. Enlarging with amino acids the range of possible abiotic organic compounds formed in the terrestrial crust offers an additional opportunity for (organo)heterotrophy to operate in these environments⁴⁸. Anaerobic amino acid fermentation, known as the Stickland reaction, involves amino acids, possibly of the same type, serving both as electron donors and acceptors⁴⁹. The Stickland reaction is a typical pathway of anaerobic bacteria belonging to the Firmicutes phylum, which can be found in serpentinization environments in which electron acceptors are lacking⁴⁸. Whether abiotic amino acids may represent valuable substrates for ecosystems inhabiting serpentinites, and whether they have shaped ancient microbial metabolisms or had a role in the emergence of a first form of biochemistry on Earth remains to be addressed. Nevertheless, the formation pathway proposed here nurtures the hydrothermal origin of life debate with an attractive alternative to the commonly considered Strecker and Fischer–Tropsch-type reactions.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability are available at <https://doi.org/10.1038/s41586-018-0684-z>.

Received: 14 February 2018; Accepted: 12 October 2018;

Published online 7 November 2018.

- McCollom, T. M. & Seewald, J. S. Abiotic synthesis of organic compounds in deep-sea hydrothermal environments. *Chem. Rev.* **107**, 382–401 (2007).
- Mével, C. Serpentinization of abyssal peridotites at mid-ocean ridges. *C. R. Geosci.* **335**, 825–852 (2003).
- Martin, W., Baross, J., Kelley, D. & Russell, M. J. Hydrothermal vents and the origin of life. *Nat. Rev. Microbiol.* **6**, 805–814 (2008).
- Sleep, N. H., Meibom, A., Fridriksson, T., Coleman, R. G. & Bird, D. K. H₂-rich fluids from serpentinization: geochemical and biotic implications. *Proc. Natl Acad. Sci. USA* **101**, 12818–12823 (2004).
- Kelley, D. S. et al. An off-axis hydrothermal vent field near the Mid-Atlantic Ridge at 30° N. *Nature* **412**, 145–149 (2001).
- Russell, M. J. The alkaline solution to the emergence of life: energy, entropy and early evolution. *Acta Biotheor.* **55**, 133–179 (2007).
- Konn, C., Charlou, J. L., Holm, N. G. & Mousis, O. The production of methane, hydrogen, and organic compounds in ultramafic-hosted hydrothermal vents of the Mid-Atlantic Ridge. *Astrobiology* **15**, 381–399 (2015).
- Lang, S. Q., Früh-Green, G. L., Bernasconi, S. M. & Butterfield, D. A. Sources of organic nitrogen at the serpentinite-hosted Lost City hydrothermal field. *Geobiology* **11**, 154–169 (2013).
- Pizzarello, S., Williams, L. B., Lehman, J., Holland, G. P. & Yarger, J. L. Abundant ammonia in primitive asteroids and the case for a possible exobiology. *Proc. Natl Acad. Sci. USA* **108**, 4303–4306 (2011).
- Elsila, J. E. et al. Meteoritic amino acids: diversity in compositions reflects parent body histories. *ACS Cent. Sci.* **2**, 370–379 (2016).
- Elmaleh, A. et al. Formation and transformations of Fe-rich serpentines by asteroidal aqueous alteration processes: a nanoscale study of the Murray chondrite. *Geochim. Cosmochim. Acta* **158**, 162–178 (2015).
- Blackman, D. K. et al. Drilling constraints on lithospheric accretion and evolution at Atlantis Massif, Mid-Atlantic Ridge 30° N. *J. Geophys. Res.* **116**, B07103 (2011).

13. Delacour, A., Früh-Green, G. L., Bernasconi, S. M., Schaeffer, P. & Kelley, D. S. Carbon geochemistry of serpentinites in the Lost City Hydrothermal System (30°N, MAR). *Geochim. Cosmochim. Acta* **72**, 3681–3702 (2008).
14. Bisio, C. et al. Understanding physico-chemical properties of saponite synthetic clays. *Microporous Mesoporous Mater.* **107**, 90–101 (2008).
15. Pisapia, C., Jamme, F., Duponchel, L. & Ménez, B. Tracking hidden organic carbon in rocks using chemometrics and hyperspectral imaging. *Sci. Rep.* **8**, 2396 (2018).
16. Klein, F. et al. Magnetite in seafloor serpentinite—some like it hot. *Geology* **42**, 135–138 (2014).
17. Nozaka, T., Fryer, P. & Andreani, M. Formation of clay minerals and exhumation of lower-crustal rocks at Atlantis Massif, Mid-Atlantic Ridge. *Geochim. Geophys. Geosyst.* **9**, Q11005 (2008).
18. Determann, S., Lobbes, J. M., Reuter, R. & Rullkötter, J. Ultraviolet fluorescence excitation and emission spectroscopy of marine algae and bacteria. *Mar. Chem.* **62**, 137–156 (1998).
19. Kumamoto, Y., Fujita, K., Smith, N. I. & Kawata, S. Deep-UV biological imaging by lanthanide ion molecular protection. *Biomed. Opt. Express* **7**, 158–170 (2015).
20. Pavlov, N. et al. Asymmetric synthesis of β^2 -tryptophan analogues via Friedel–Crafts alkylation of indoles with a chiral nitroacrylate. *J. Org. Chem.* **76**, 6116–6124 (2011).
21. Jamme, F. et al. Synchrotron UV fluorescence microscopy uncovers new probes in cells and tissues. *Microsc. Microanal.* **16**, 507–514 (2010).
22. Sanni, O. D., Wagner, M. S., Briggs, D., Castner, D. G. & Vickerman, J. C. Classification of adsorbed protein static ToF-SIMS spectra by principal component analysis and neural networks. *Surf. Interface Anal.* **33**, 715–728 (2002).
23. Steele, A., Toporski, J. K. W., Avci, R., Guidry, S. & McKay, D. S. Time of flight secondary ion mass spectrometry (ToF-SIMS) of a number of hopanoids. *Org. Geochem.* **32**, 905–911 (2001).
24. Toporski, J. K. W. & Steele, A. Characterization of purified biomarker compounds using time of flight-secondary ion mass spectrometry (ToF-SIMS). *Org. Geochem.* **35**, 793–811 (2004).
25. Siljeström, S. et al. Detection of organic biomarkers in crude oils using ToF-SIMS. *Org. Geochem.* **40**, 135–143 (2009).
26. Chiriboga, L. et al. Infrared spectroscopy of human tissue. I. Differentiation and maturation of epithelial cells in the human cervix. *Biospectroscopy* **4**, 47–53 (1998).
27. Ménez, B., Pasini, V. & Brunelli, D. Life in the hydrated suboceanic mantle. *Nat. Geosci.* **5**, 133–137 (2012).
28. Kooli, F. & Jones, W. Characterization and catalytic properties of a saponite clay modified by acid activation. *Clay Miner.* **32**, 633–643 (1997).
29. Molina, C. B., Casas, J. A., Pizarro, A. H. & Rodríguez, J. J. in *Clay: Types, Properties and Uses* (eds Humphrey, J. P. & Boyd, D. E.) 435–474 (Nova Science Publisher, New York, 2011).
30. Williams, L. B. et al. in *Earliest Life on Earth: Habitats, Environments and Methods of Detection* (eds Golding, S. D. & Gliksun, M.) 79–112 (Springer, Amsterdam, 2010).
31. Meunier, A., Petit, S., Cockell, C. S., El Albani, A. & Beaufort, D. The Fe-rich clay microsystems in basalt-komatiite lavas: importance of Fe-smectites for pre-biotic molecule catalysis during the Hadean eon. *Orig. Life Evol. Biosph.* **40**, 253–272 (2010).
32. Belver, C., Bañares-Muñoz, M. A. & Vicente, M. A. Fe-saponite pillared and impregnated catalysts: I. Preparation and characterization. *Appl. Catal. B* **50**, 101–112 (2004).
33. Choudary, B. M., Kantam, M. L., Sateesh, M., Rao, K. K. & Santhi, P. L. Iron pillared clays — efficient catalysts for Friedel–Crafts reactions. *Appl. Catal. A* **149**, 257–264 (1997).
34. Rueping, M. & Nachtsheim, B. J. A review of new developments in the Friedel–Crafts alkylation — from green chemistry to asymmetric catalysis. *Beilstein J. Org. Chem.* **6**, 6 (2010).
35. Milesi, V., McCollom, T. M. & Guyot, F. Thermodynamic constraints on the formation of condensed carbon from serpentinization fluids. *Geochim. Cosmochim. Acta* **189**, 391–403 (2016).
36. Cody, G. D. et al. Primordial carbonylated iron-sulfur compounds and the synthesis of pyruvate. *Science* **289**, 1337–1340 (2000).
37. Seyfried, W. E. Jr, Pester, N. J., Tutolo, B. M. & Ding, K. The Lost City hydrothermal system: constraints imposed by vent fluid chemistry and reaction path models on subsurface heat and mass transfer processes. *Geochim. Cosmochim. Acta* **163**, 59–79 (2015).
38. Proskurowski, G., Lilley, M. D., Kelley, D. S. & Olson, E. J. Low temperature volatile production at the Lost City Hydrothermal Field, evidence from a hydrogen stable isotope geothermometer. *Chem. Geol.* **229**, 331–343 (2006).
39. Brandes, J. A. et al. Abiotic nitrogen reduction on the early Earth. *Nature* **395**, 365–367 (1998).
40. Schoonen, M. A. & Xu, Y. Nitrogen reduction under hydrothermal vent conditions: implications for the prebiotic synthesis of C-H-O-N compounds. *Astrobiology* **1**, 133–142 (2001).
41. Salmon, V., Derenne, S., Lallier-Vergès, E., Largeau, C. & Beaudoin, B. Protection of organic matter by mineral matrix in a Cenomanian black shale. *Org. Geochem.* **31**, 463–474 (2000).
42. Pearson, V. K. et al. Clay mineral–organic matter relationships in the early solar system. *Meteorit. Planet. Sci.* **37**, 1829–1833 (2002).
43. Manuella, F. C., Carbone, S. & Barreca, G. Origin of saponite-rich clays in a fossil serpentinite-hosted hydrothermal system in the crustal basement of the Hyblean Plateau (Sicily, Italy). *Clays Clay Miner.* **60**, 18–31 (2012).
44. Arndt, N. T. & Nisbet, E. G. Processes on the young Earth and the habitats of early life. *Annu. Rev. Earth Planet. Sci.* **40**, 521–549 (2012).
45. Granold, M., Hajieva, P., Toşa, M. I., Irimie, F.-D. & Moosmann, B. Modern diversification of the amino acid repertoire driven by oxygen. *Proc. Natl Acad. Sci. USA* **115**, 41–46 (2018).
46. Ruiz-Mirazo, K., Briones, C. & de la Escosura, A. Prebiotic systems chemistry: new perspectives for the origins of life. *Chem. Rev.* **114**, 285–366 (2014).
47. Russell, M. J., Daniel, R. M., Hall, A. J. & Sherrington, J. A. A hydrothermally precipitated catalytic iron sulphide membrane as a first step toward life. *J. Mol. Evol.* **39**, 231–243 (1994).
48. Schrenk, M. O., Brazelton, W. J. & Lang, S. Q. in *Carbon in Earth* (eds Hazen, R. M. et al.) **75**, 575–606 (Mineralogical Society of America, Chantilly, 2013).
49. Barker, H. A. in *The Bacteria. A Treatise on Structure and Function* (eds Gunsalus, I. C. & Stanier, R. Y.) 151–207 (Academic, Cambridge, 1961).

Acknowledgements We thank B. Van de Moortèle for the FIB sections, O. Boudouma for assistance during SEM experiments and V. Pasini, D. Brunelli, M. Chaussidon and J. Badro for help and discussion. We acknowledge the IODP program (<https://www.iodp.org/>) and SOLEIL synchrotron for granted access to DISCO and SMIS beamlines. This research was supported by the Deep Carbon Observatory, the deepOASES ANR project (ANR-14-CE01-0008) and the French CNRS (Défi Origines M.I. 2018). This is IGP contribution no. 3976.

Reviewer information Nature thanks J. Baross, M. Russell and the anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions B.M., C.P. and M.A. conceived the research. B.M., C.P., F.J., M.R., Q.P.V., A.B., M.A. and P.D. performed the experiments. L.R. performed the thermodynamic calculations. B.M. wrote the manuscript. All authors contributed to the interpretation of the data and commented on the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0684-z>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0684-z>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to B.M. or M.A.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size.

Sample preparation. Drilled rock samples were sawed with sterile ultrapure water to extract the inner core, free of possible post-sampling contamination. The saw was previously treated with 5% sodium hypochlorite, and then rinsed twice with sterile ultrapure water. The inner core was then manipulated using clean pliers, thinned and polished manually on both faces (down to a thickness of tens of micrometres) with pure ethanol using silicon carbide polishing disks without any use of resin or glue. Samples were then sequentially analysed using, by order of potential beam-induced damages and sample preparation constraints (for example, metallic coating or sample milling), S-FTIR and S-DUV microspectroscopy, TOF-SIMS imaging, SEM and TEM.

S-FTIR. S-FTIR hyperspectral imaging was performed at the SMIS beamline (SOLEIL synchrotron radiation facility, Saint Aubin, France)⁵⁰ by taking advantage of the brightness of the bending magnet radiation and of the confocal geometry of the microscope objective⁵¹. The sensitivity of infrared microspectroscopy is on average 10^{-4} M (10^{-12} g of molecules)⁵⁰. Data were acquired in transmission mode with a Nicplan microscope coupled to a Magna 550 FT-IR spectrometer (Thermo-Nicolet). The confocal aperture was set at $5 \times 5 \mu\text{m}^2$ using a $32\times$ infinity corrected Schwarzschild objective with a numerical aperture of 0.65 (Nicolet Refflachromat) and a matching $10\times$ condenser. The microscope is equipped with a motorized sample stage (repeatability $1 \mu\text{m}$) and a liquid-nitrogen-cooled Mercury Cadmium Telluride detector (MCT-A; detector element size $250 \mu\text{m}$). The sample was deposited on a CaF_2 window without any treatment. Hyperspectral data cubes in which pixels correspond to individual S-FTIR spectra were collected in the $4,000\text{--}800 \text{ cm}^{-1}$ mid-infrared range using a step size of $5 \mu\text{m}$. Acquisitions were carried out using 50 accumulations per spectrum/pixel with a spectral resolution set at 4 cm^{-1} . Spectrum analyses were first performed using OMNIC software (Thermo Fisher Scientific) to obtain distribution maps of the aliphatic CH_2/CH_3 stretching band area between $2,800$ and $3,000 \text{ cm}^{-1}$ and then using an approach combining principal component analysis and multivariate curve resolution–alternating least-squares analysis implemented in the MATLAB software and the PLS toolbox (Eigenvector Research)¹⁵.

S-DUV. S-DUV microspectroscopy enables imaging with a nanomolar sensitivity of aromatic, phenolic or unsaturated compounds without any external fluorescent probes²¹. Fluorescence imaging was carried out at the DISCO beamline (SOLEIL synchrotron radiation facility, Saint Aubin, France), where two complementary full UV compatible microscopes are coupled to the monochromatized synchrotron radiation continuously emitted from a bending magnet, which allows fluorescence excitation down to 180 nm (from 250 to 310 nm in the present study) at a sub-micrometric spatial resolution⁵². The use of a synchrotron light as a UV source enables the excitation light to be precisely tuned to the absorption of endogenous fluorochromes. Both microscopes, equipped with motorized sample stages (repeatability $1 \mu\text{m}$), were used with a Zeiss Ultrafluor $40\times$ (glycerine immersion) objective. The sample was deposited on a quartz window without any treatment. To localize fluorescent areas within serpentinites, a full-field Zeiss Axio Observer Z-1 inverted microscope was first used. The fluorescent signal was collected by a PIXIS 1024 BUUV camera (Princeton Instruments) with bandpass filters at $327\text{--}353$, $370\text{--}410$, $412\text{--}438$ and $400\text{--}440 \text{ nm}$ (Semrock) and associated integration times of 120 s . Images were analysed using the Fiji software⁵³ and stitched by linear blending⁵⁴. Microspectrofluorescence emission spectra in the range of $285\text{--}550 \text{ nm}$ were thereafter collected with a $70\% \text{ C}$ Peltier-cooled iDus charge-coupled device (CCD) detector (Andor) of $1,024 \times 256$ pixels with a $26 \times 26 \mu\text{m}^2$ pixel size on selected areas using a spectral Olympus IX71 inverted microscope. Hyperspectral data cubes in which pixels correspond to individual fluorescence spectra were collected from areas up to $100 \times 80 \mu\text{m}^2$ in size with a $2\text{--}3 \mu\text{m}$ step size and $20\text{--}40\text{-s}$ acquisition time per spectrum/pixel. With the exception of manual removal of spikes coming from cosmic rays, no filtering or treatment of the autofluorescence spectra was conducted. Deconvoluted images of each individual fluorescent component were then produced with the Labspec software (Jobin-Yvon) using Gaussian functions and 10 iterations.

TOF-SIMS. TOF-SIMS imaging allows the simultaneous detection of inorganic and organic molecules on solid surfaces without extraction, chemical preparation or derivatization. Experiments were conducted using a TOF-SIMS IV reflectron-type mass spectrometer (IONTOF GmbH) located at the Institut de Chimie des Substances Naturelles (CNRS, Gif-sur-Yvette, France)⁵⁵. The instrument is equipped with a bismuth liquid metal ion gun delivering a pulsed Bi_3^+ cluster ion beam. 25 keV primary ions impacted the sample at an incidence angle of 45° and a pulsed current of 0.21 pA . Emitted secondary ions were accelerated to 2 keV (2 kV extraction) towards a field-free region and a single stage reflectron (first-order compensation). Secondary ions ejected from the few upper monolayers of the sample surface were post-accelerated to 10 keV before reaching the detector made of a micro-channel plate, a scintillator and a photomultiplier. The ion column

focusing mode ensured a spatial resolution of $2\text{--}5 \mu\text{m}$ and a mass resolution of $5,000$ (full-width half-maximum) at m/z of 500 . A low-energy ($\sim 20 \text{ eV}$) electron flood gun was used between two successive primary ion pulses for charge compensation with minimum damage on the sample surface. An optical camera located in the sample vacuum chamber aided location of the samples, which were deposited on the sample holder without any treatment or adhesive. Ion images in both negative and positive polarities were acquired in a raster pattern on areas of $500 \times 500 \mu\text{m}^2$ and 256×256 pixels, giving a pixel size of $1.95 \times 1.95 \mu\text{m}^2$. The images were recorded with a primary ion fluence of $3.44 \times 10^{11} \text{ ions cm}^{-2}$ (100 scans with cycle time of $100 \mu\text{s}$). The accumulated primary ion dose never exceeded $10^{12} \text{ ions cm}^{-2}$, which is below the static limit for organic molecules⁵⁶. Data acquisition and processing were done using the SurfaceLab 6 software (IONTOF GmbH). Spectra from the total analysis area or from selected regions of interest were extracted. Internal mass calibration was performed using the low mass fragment ion signals of H^+ , H_2^+ , H_3^+ , C^+ , CH_3^+ , and C_2H_3^+ for the positive ion mode and C^- , CH^- , CH_2^- , C_2^- , C_3^- , and C_4H^- for the negative ion mode. Assignments of ion peaks were made according to the instrument resolution, accuracy and the valence rule. The presence of tryptophan was investigated through the identification and assignment of all required fragment ions that are characteristic of this amino acid^{57–59}. The absence of interferences was carefully checked for each assigned peak and mass deviations relatively to the theoretical m/z values were calculated. They all fall within admitted values for TOF-SIMS analyses (Supplementary Table 2). Image reconstruction for selected fragment ions was carried out by integrating signal intensities at given m/z values across the dataset.

SEM. SEM was performed at the Service Commun de Microscopie Electronique à Balayage (UPMC, Paris, France) on Au-coated samples with a Zeiss SUPRA 55 VP field emission microscope operating at 3 to 15 kV accelerating voltage at respectively low and high currents (from 10 pA to 1 nA). Images were collected using secondary electron (SE) detectors (Everhart-Thornley and InLens for high- and low-voltage mode, respectively) and a backscattered electron (BSE) detector (AsB). Images were further processed with the ImageJ software⁶⁰ for contrast and brightness adjustment.

FIB milling and TEM. As TEM requires samples to be electron transparent, ultrathin foils (thickness $< 100 \text{ nm}$) were milled using a Zeiss NVision 40 cross beam microscope (CLYM, University of Lyon, France) which combines a high-resolution field emission SEM with a Seiko FIB column. The $10\text{--}15\text{-nm}$ -thick Au layer previously deposited for SEM observations prevented from amorphization of the subsurface. In addition, before excavation, milled volumes were protected by a FIB-induced $1\text{--}2 \mu\text{m}$ -thick carbon coating (Extended Data Figs. 4a–c, 7a–c). To allow cross-sectional observations at depth (Extended Data Figs. 4c–e, 7c, d), excavations were first made on one side of the TEM foil location using a Ga^+ beam, emitted by a Ga liquid metal ion source operating at 30 kV accelerating voltage with ion beam current of decreasing intensities (13 , 6.5 , 3 nA and 700 pA). Cross-sectional images were collected at 5 kV using a secondary electrons detector (SESI) and InLens detector for high and low voltage mode, respectively, and at 1.25 kV using an energy selective backscattered (BSE) detector. Elemental analyses were carried out with an Oxford Instrument energy dispersive X-ray spectrometer (EDS) (X-max 50 mm^2 silicon drift detector). TEM sections were then extracted from the bulk sample following excavation of its second side once a thickness of $\sim 1\text{--}2 \mu\text{m}$ was reached. TEM sections were then fixed by C ion beam deposition on a half copper TEM grid. Further thinning of the TEM foil to few tens of nanometres was obtained with a glancing angle beam at low ion beam current (from 700 pA to 50 pA at 30 kV). They were finally cleaned for traces of Ga ion implantation by a milling at 2 kV and 50 pA during $3\text{--}5 \text{ min}$ on each side. TEM observations were then carried out with a TOPCON microscope operating at 200 kV (CLYM, University of Lyon, France). Images, further processed with ImageJ software⁶⁰ for contrast and brightness adjustment, have been collected at a nanometric spatial resolution with a CCD Camera.

Electron microprobe analysis. Electron microprobe analyses were acquired in punctual mode on carbon-coated petrographic thin sections using the Cameca SX100 installed at Geosciences Montpellier (France). Operating conditions were 20 keV and 10 nA .

Thermodynamic constraints on the abiotic synthesis of tryptophan at Lost City hydrothermal field. To evaluate the extent to which the temperature-pressure-composition conditions prevailing in the Lost City hydrothermal system are conducive to the abiotic synthesis of tryptophan and some of its possible precursors (indole and pyruvate) (Extended Data Fig. 9), the chemical affinities A_r of their formation reactions were computed at 100°C from the relation

$$A_r = RT \ln(K_r/Q_r)$$

in which R and T denote the gas constant and absolute temperature, respectively; Q_r is the reaction quotient; and K_r is the thermodynamic equilibrium constant. The reaction quotient is calculated from the relation

$$Q_r = \prod_i a_i^{n_{i,r}}$$

in which a_i represents the activity of the i th species involved in the r th reaction, and $n_{i,r}$ is the stoichiometric coefficient of the species in the reaction. Note that the chemical affinities computed below are mathematically equivalent to the opposite of the overall Gibbs energies ΔG_r computed previously⁶¹ for the abiotic synthesis of amino acids in accord with

$$\Delta G_r = \Delta G_r^\circ + RT \ln Q_r$$

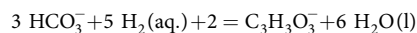
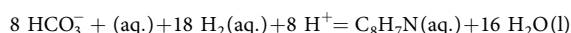
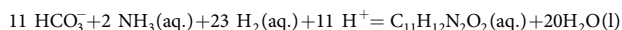
where ΔG_r° is the standard Gibbs energy of the abiotic synthesis reaction.

In the absence of Helgeson–Kirkham–Flowers (HKF) parameters for aqueous indole and pyruvate, the values of the equilibrium constants at 100 °C were estimated using the van't Hoff equation

$$\log K_{373.15\text{K}} = \log K_{298.15\text{K}} + \frac{\Delta H_r^\circ}{RT} \left(\frac{1}{373.15} - \frac{1}{298.15} \right)$$

in which ΔH_r° is the standard enthalpy of the reaction at 298.15 K. The equilibrium constants at 25 °C and 100 °C have been carried out using standard Gibbs energies and enthalpies of formation as described previously for the aqueous inorganic species^{62,63} and for aqueous tryptophan, indole and pyruvate⁶⁴.

Assuming a pH of 8.5 for a hypothetical Lost City fluid cooled to 100 °C during its ascent towards the sea floor³⁷, the predominant inorganic carbon and nitrogen species are HCO_3^- and $\text{NH}_3(\text{aq})$ (Extended Data Fig. 10). Accordingly, the abiotic synthesis reactions have been written as



in which $\text{C}_{11}\text{H}_{12}\text{N}_2\text{O}_2$ denotes tryptophan, $\text{C}_8\text{H}_7\text{N}$ denotes indole, and $\text{C}_3\text{H}_3\text{O}_3^-$ denotes pyruvate. The reaction quotients were evaluated with activities for the aqueous (aq.) species that approximate the concentrations reported for the Lost City hydrothermal fluids; that is, $a_{\text{CO}_2(\text{aq.})} = 10^{-2.54}$ (ref. ⁶⁵), $a_{\text{H}_2(\text{aq.})} = 10^{-2}$ (ref. ³⁸), and $a_{\text{NH}_3(\text{aq.})} = 10^{-6}$ (ref. ⁸). Nanomolar concentrations have been assumed for the organic compounds.

The chemical affinities calculated for the abiotic synthesis of tryptophan, indole and pyruvate are, respectively, 134.5 kJ mol⁻¹, 193.48 kJ mol⁻¹ and 43.07 kJ mol⁻¹. The chemical affinity for the synthesis of tryptophan is notably more favourable than that calculated previously⁶¹ because of the higher concentration of H_2 in the Lost City hydrothermal fluids. The dependence of the chemical affinities of the reactions on the activity of aqueous H_2 can be deduced from Extended Data Fig. 9.

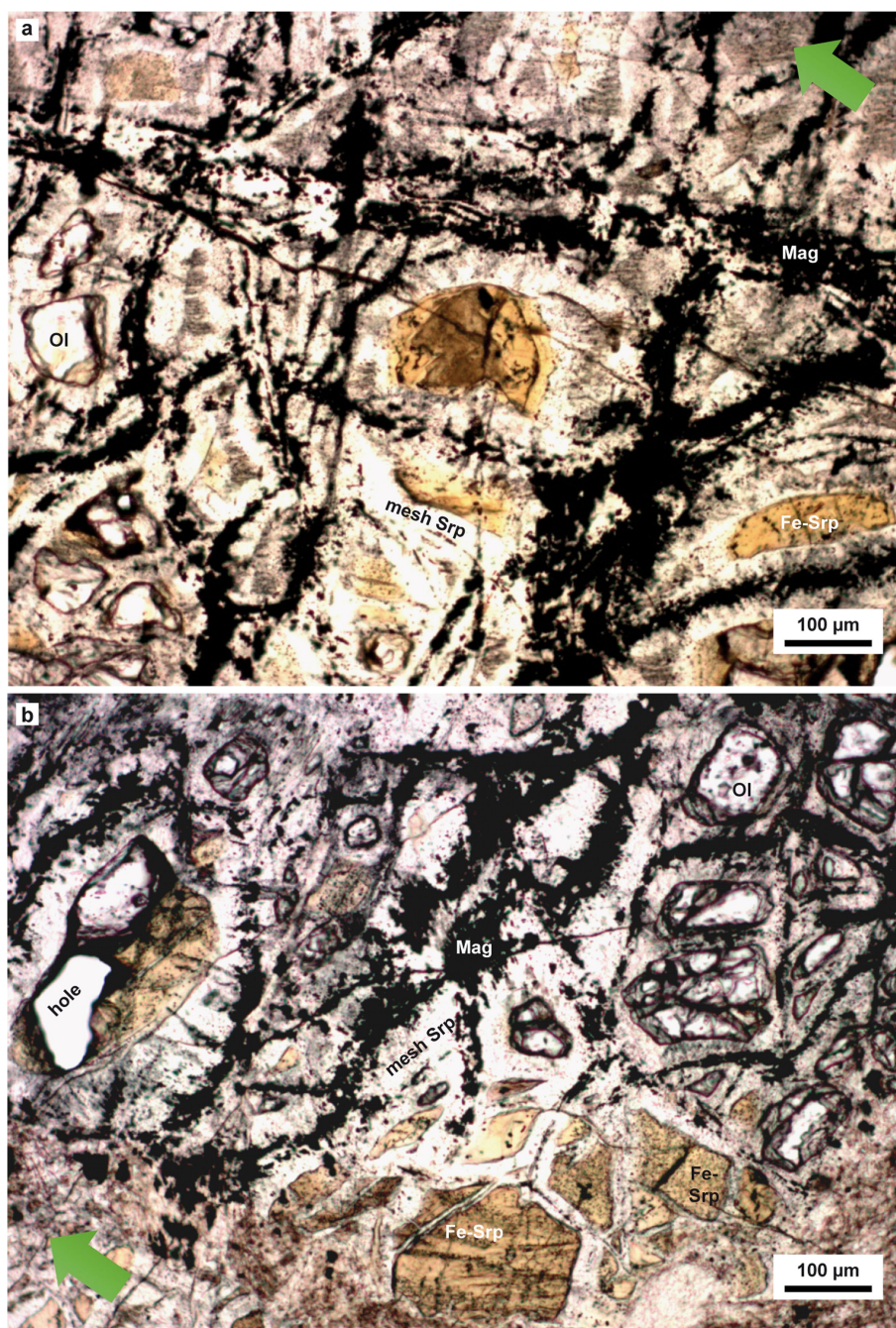
It can be seen that the abiotic synthesis of tryptophan becomes favourable only for H_2 concentrations above approximately $10^{-2.8} \text{ m}$.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

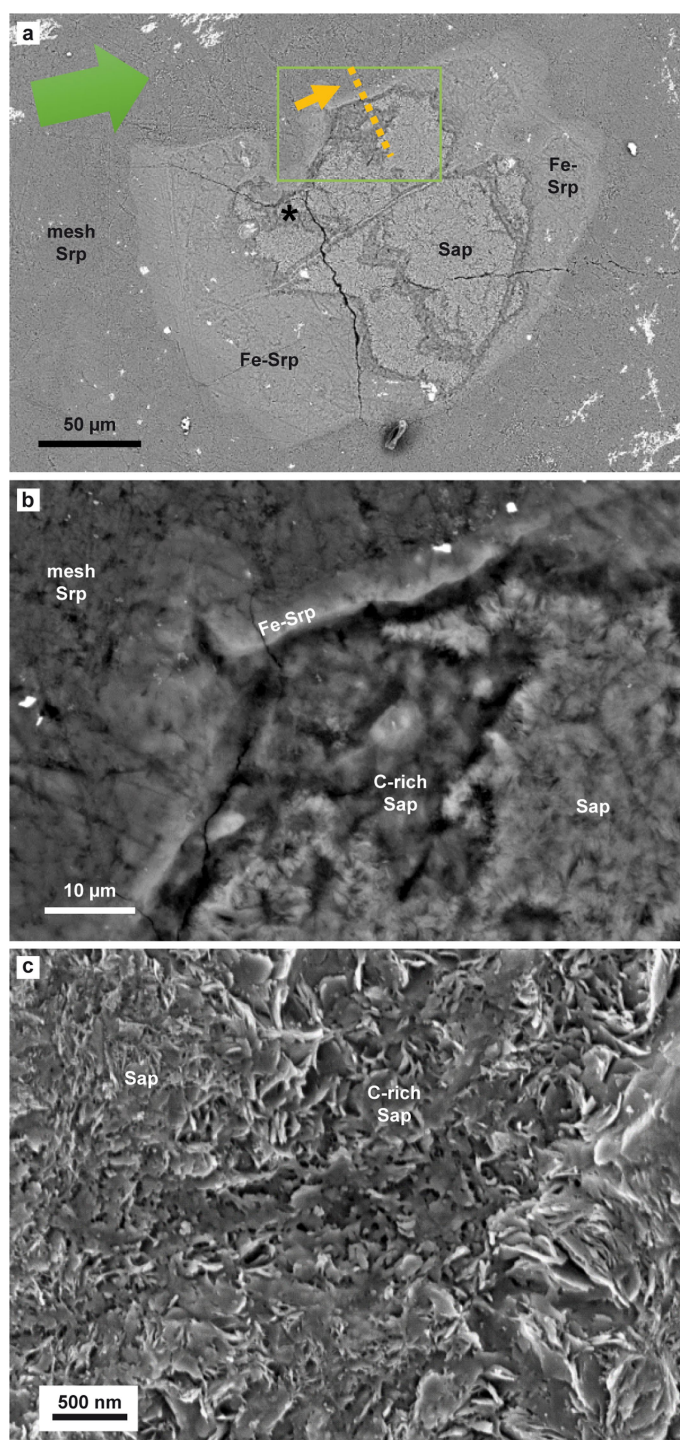
The data supporting the findings of this study are available within the paper, its Extended Data and its Supplementary Information. The datasets generated and analysed in this study are available from the corresponding author upon reasonable request.

50. Dumas, P. et al. Synchrotron infrared microscopy at the French Synchrotron Facility SOLEIL. *Infrared Phys. Technol.* **49**, 152–160 (2006).
51. Jamme, F., Lagarde, B., Giuliani, A., Garcia, G. A. & Mercury, L. Synchrotron infrared confocal microscope: application to infrared 3D spectral imaging. *J. Phys. Conf. Ser.* **425**, 142002 (2013).
52. Giuliani, A. et al. DISCO: a low-energy multipurpose beamline at synchrotron SOLEIL. *J. Synchrotron Radiat.* **16**, 835–841 (2009).
53. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
54. Preibisch, S., Saalfeld, S. & Tomancak, P. Globally optimal stitching of tiled 3D microscopic image acquisitions. *Bioinformatics* **25**, 1463–1465 (2009).
55. Brunelle, A., Touboul, D. & Laprévotte, O. Biological tissue imaging with time-of-flight secondary ion mass spectrometry and cluster ion sources. *J. Mass Spectrom.* **40**, 985–999 (2005).
56. Touboul, D., Kollmer, F., Niehuis, E., Brunelle, A. & Laprévotte, O. Improvement of biological time-of-flight-secondary ion mass spectrometry imaging with a bismuth cluster ion source. *J. Am. Soc. Mass Spectrom.* **16**, 1608–1618 (2005).
57. Mazel, V. et al. Identification of ritual blood in African artifacts using TOF-SIMS and synchrotron radiation microspectroscopies. *Anal. Chem.* **79**, 9253–9260 (2007).
58. Cersoy, S., Richardin, P., Walter, P. & Brunelle, A. Cluster TOF-SIMS imaging of human skin remains: analysis of a South-Andean mummy sample. *J. Mass Spectrom.* **47**, 338–346 (2012).
59. Farre, B. et al. Shell layers of the black-lip pearl oyster *Pinctada margaritifera*: matching microstructure and composition. *Comp. Biochem. Physiol. B* **159**, 131–139 (2011).
60. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
61. Amend, J. P. & Shock, E. L. Energetics of amino acid synthesis in hydrothermal ecosystems. *Science* **281**, 1659–1662 (1998).
62. Shock, E. L. & Helgeson, H. C. Calculation of the thermodynamic and transport properties of aqueous species at high pressures and temperatures: correlation algorithms for ionic species and equation of state predictions to 5 kb and 1000 °C. *Geochim. Cosmochim. Acta* **52**, 2009–2036 (1988).
63. Shock, E. L., Helgeson, H. C. & Sverjensky, D. A. Calculation of the thermodynamic and transport properties of aqueous species at high pressures and temperatures: standard partial molal properties of inorganic neutral species. *Geochim. Cosmochim. Acta* **53**, 2157–2183 (1989).
64. Tewari, Y. B. & Goldberg, R. N. An equilibrium and calorimetric investigation of the hydrolysis of L-tryptophan to (indole + pyruvate + ammonia). *J. Solution Chem.* **23**, 167–184 (1994).
65. Proskurowski, G. et al. Abiogenic hydrocarbon production at Lost City hydrothermal field. *Science* **319**, 604–607 (2008).
66. Bakke, Ø. & Mostad, A. The structure and conformation of tryptophan in the crystal of the pure racemic compound and the hydrogen oxalate. *Acta Chem. Scand. B* **34**, 559–570 (1980).



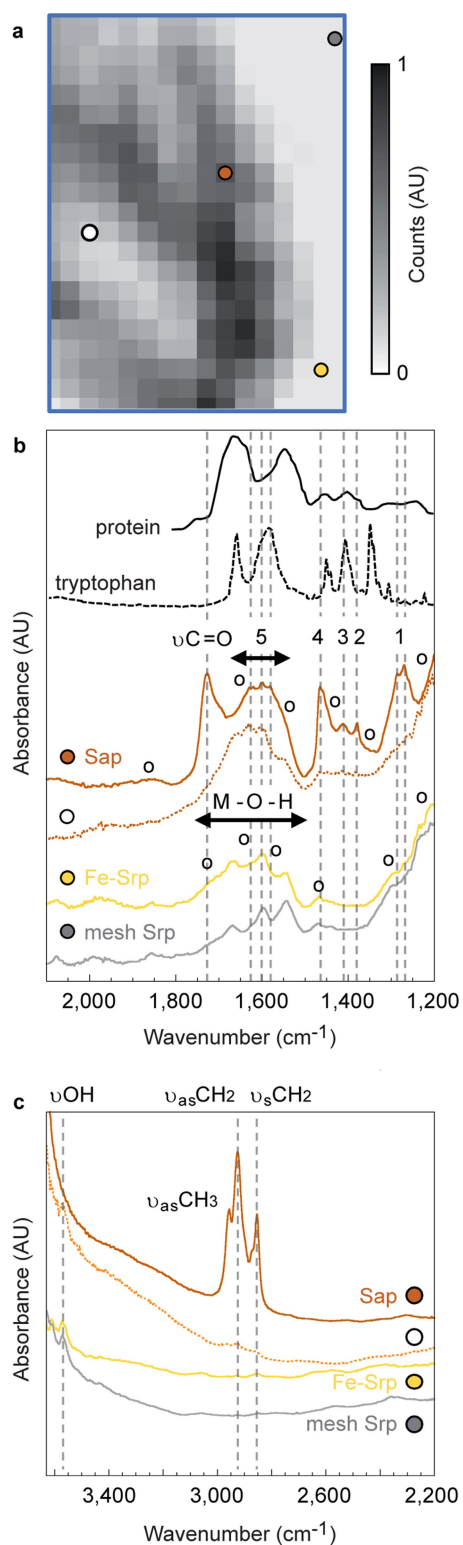
Extended Data Fig. 1 | Large optical views of the deeply serpentinized harzburgite recovered by drilling the Atlantis Massif (173.15 m below sea floor) during the IODP Expedition 304 at Hole U1309D¹². **a, b,** Both photomicrographs are centred on the two areas described in the present study (in Figs. 1–3, Extended Data Figs. 2–5 for **a** and Extended Data Figs. 6, 7 for **b**, respectively). The high-temperature hydrated paragenesis is composed of serpentine and magnetite, both after olivine (Ol), and

forming a characteristic mesh texture. Yellow to brownish phases are frequently found in the core of the mesh serpentine. They correspond to Fe-rich serpentine and Fe-rich saponite, formed at lower temperature during secondary and tertiary alteration reactions that occur at the expense of the mesh serpentine and olivine kernels¹⁵, some remnants of which can still be observed. Hole figures an olivine crystal removed during sample thinning and polishing. Green arrows indicate sample orientation.



Extended Data Fig. 2 | SEM images of the Fe-rich saponite enriched in organic carbon. **a, b**, SEM-BSE images collected at 15 kV on the mineral assemblage displayed in Figs. 1, 2 and Extended Data Fig. 1a. **b**, Magnified view of the area represented by the green box in **a**. The green arrow indicates sample orientation; the orange dashed line and the associated arrow denote the location and front face, respectively, of the FIB foil milled for TEM observations (Fig. 3a, b, Extended Data Fig. 4). Textures and differences in grey levels resulting from chemical contrasts allow the

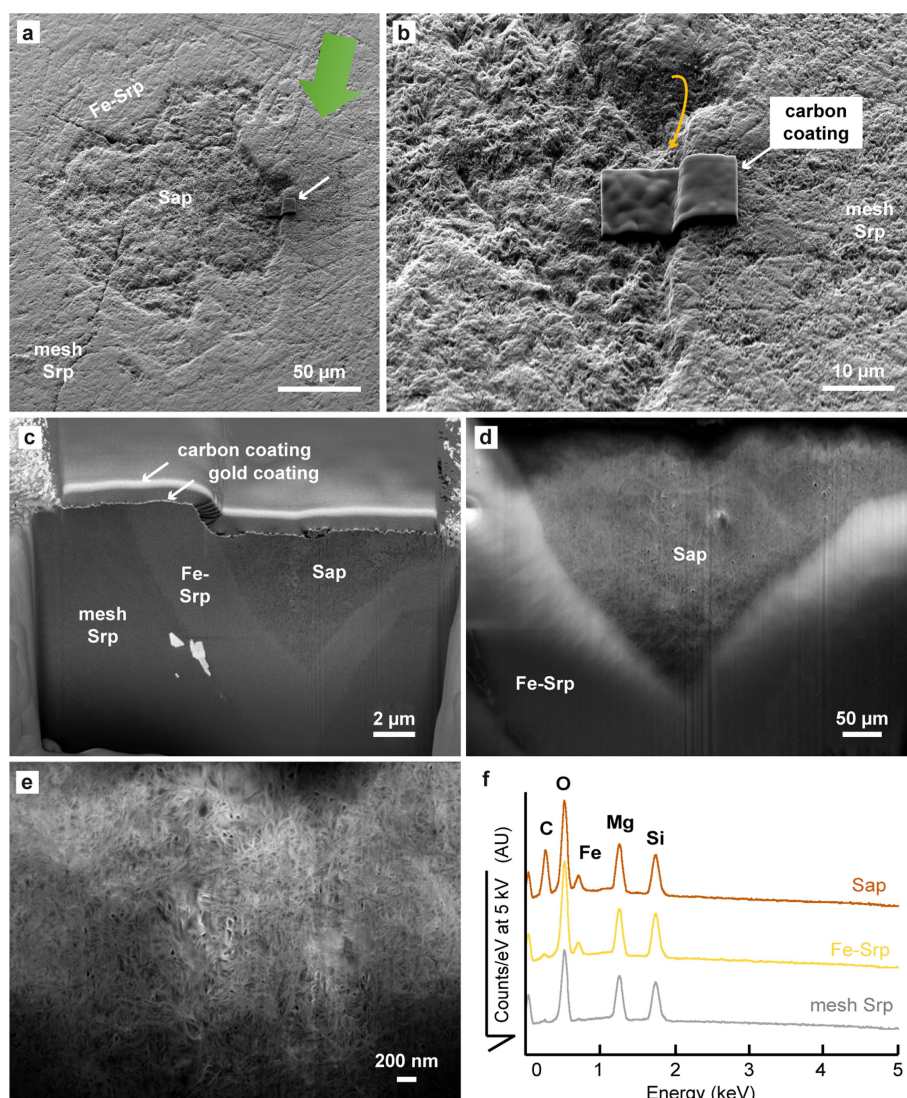
recognition of each mineral phase previously characterized by S-FTIR and electron microprobe analysis, namely mesh serpentine, Fe-rich serpentine and Fe-rich saponite. **c**, SEM-SE image collected at 3 kV accelerating voltage (location shown by an asterisk in **a**), in which the characteristic platelets of saponite¹⁴ are well recognizable. Fe-rich saponite presents distinct grey levels in **a** and **b** that relate to its variable content in organic carbon, hence darkening its aspect when carbon is abundant. The image in panel **a** was adapted from ref. ¹⁵, Springer Nature Limited.



Extended Data Fig. 3 | See next page for caption.

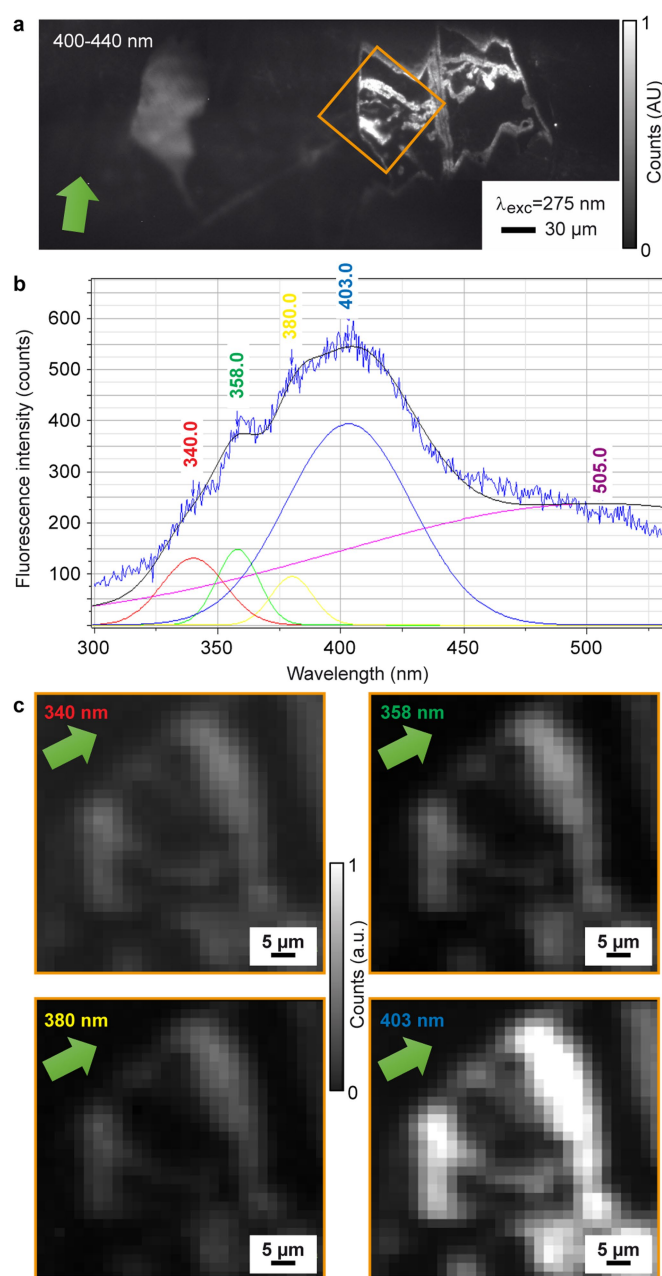
Extended Data Fig. 3 | S-FTIR confirmed the presence of N-bearing organic compounds in the Fe-rich saponite. **a**, S-FTIR distribution maps of the aliphatic CH₂/CH₃ stretching band area between 2,800 and 3,000 cm⁻¹ shown in **c** and collected in the area indicated by the blue box in Fig. 1a. **b**, **c**, Associated S-FTIR spectra. The spectrum collected in the C-rich saponite (Extended Data Fig. 2a, b) shows the presence of organic compounds with modes at (1) 1,270 cm⁻¹ and 1,285 cm⁻¹, (2) 1,380 cm⁻¹, (3) 1,412 cm⁻¹, (4) 1,460–1,465 cm⁻¹, (5) 1,550–1,650 cm⁻¹, and 1,728 cm⁻¹ in **b**, and 2,855, 2,871, 2,924 and 2,958 cm⁻¹ in **c**. Band assignments are compiled in Supplementary Table 3. Contributions of the H–O–H bending from the saponite interlayer water at 1,627 cm⁻¹ may interfere¹⁵. Also shown are the S-FTIR spectra collected in the

mesh serpentine and the Fe-rich serpentine, both being nearly depleted in absorption bands related to organic compounds. They show instead characteristic O–H stretching bands at 3,570 and 3,610 cm⁻¹ and M–O–H bending modes (with M indicating any of the cations in the hydrated silicate structure) in the range 1,500–1,680 cm⁻¹. Dotted brown curves correspond to a mixture of saponite and serpentine. Precise locations of analysis are indicated in **a** with the corresponding coloured dots. FTIR spectra of protein²⁶ and L-tryptophan (<https://webbook.nist.gov/cgi/cbook.cgi?ID=C73223&Mask=80>) are shown for comparison. ‘o’ denotes overtone-combination bands; ‘ν’ denotes stretching; ‘ν_{as}’ and ‘ν_s’ denote asymmetric and symmetric stretching, respectively.

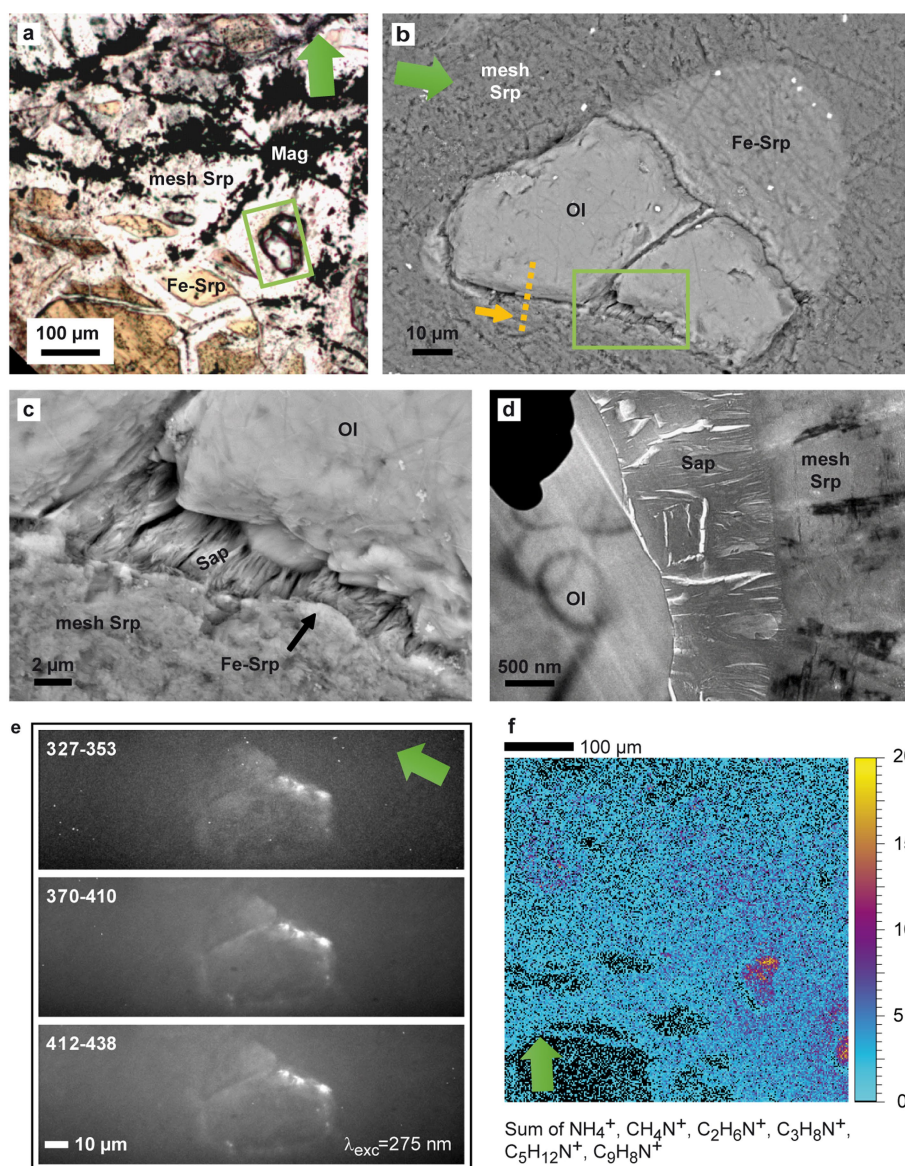


Extended Data Fig. 4 | SEM image sequence illustrating FIB milling, which allowed cross-sectional visualization of the interfaces between the C-bearing Fe-rich saponite, the Fe-rich serpentine, the mesh serpentine and the associated textures. a, SEM-SESI view at low magnification of the UV-fluorescent area depicted in Fig. 1 and Extended Data Fig. 5. The white arrow denotes the region where an ultrathin foil was milled for TEM observations (Fig. 3a, b). The green arrow provides the orientation of the sample. **b**, Enlarged SEM-SESI view of the region of interest coated with a carbon protective layer. The orange arrow denotes

the milling direction. **c**, SEM-BSE image showing the front face of the milled section. **d**, **e**, Enlarged SEM-InLens views of the Fe-rich saponite displaying a nanoporous texture with maximum pore size of less than 100 nm, in contrast to the compact Fe-rich serpentine hosted in the mesh serpentine. Pores of the Fe-rich saponite are large enough to support the presence of a 1.2-nm-sized molecule such as tryptophan⁶⁶ (see also Fig. 3a, b for associated TEM observations). **f**, Associated EDS spectra collected at 5 kV accelerating voltage.

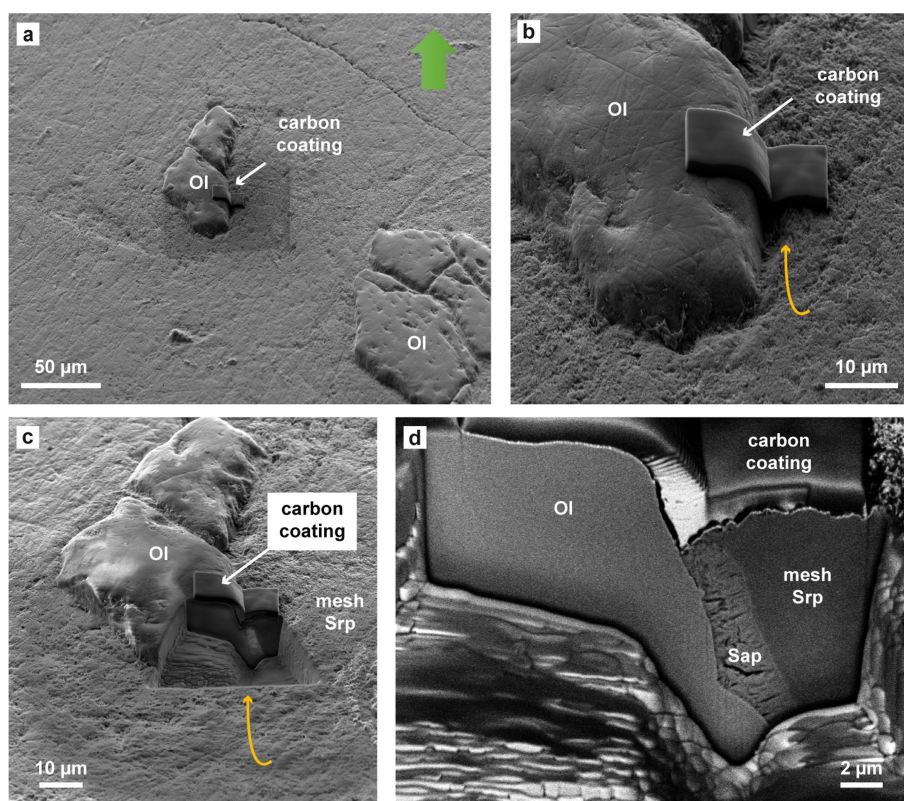


Extended Data Fig. 5 | S-DUV spectral signature of the endogenous fluorescence revealed in the Fe-rich saponite. **a**, Full-field S-DUV image displayed in Fig. 1 and collected using an excitation wavelength (λ_{exc}) of 275 nm, and a detection range of fluorescence emission between 400 and 440 nm. **b**, Fluorescence emission spectra collected with excitation wavelength of 275 nm summed from the hyperspectral datacube acquired in the area indicated by the orange box in **a** (30 s per point, 2- μ m step). Fit of the S-DUV hyperspectral maps, performed using Gaussian functions and 10 iterations, resolved 4 main contributions at 340 ± 6 , 358 ± 3 , 380 ± 3 and 403 ± 3 nm (mean \pm s.d. of three independent measurements). **c**, Associated spatial distributions of fluorescence emissions at 340, 358, 380 and 403 nm. They revealed systematic co-localization of these four components. The green arrows indicate sample orientation.



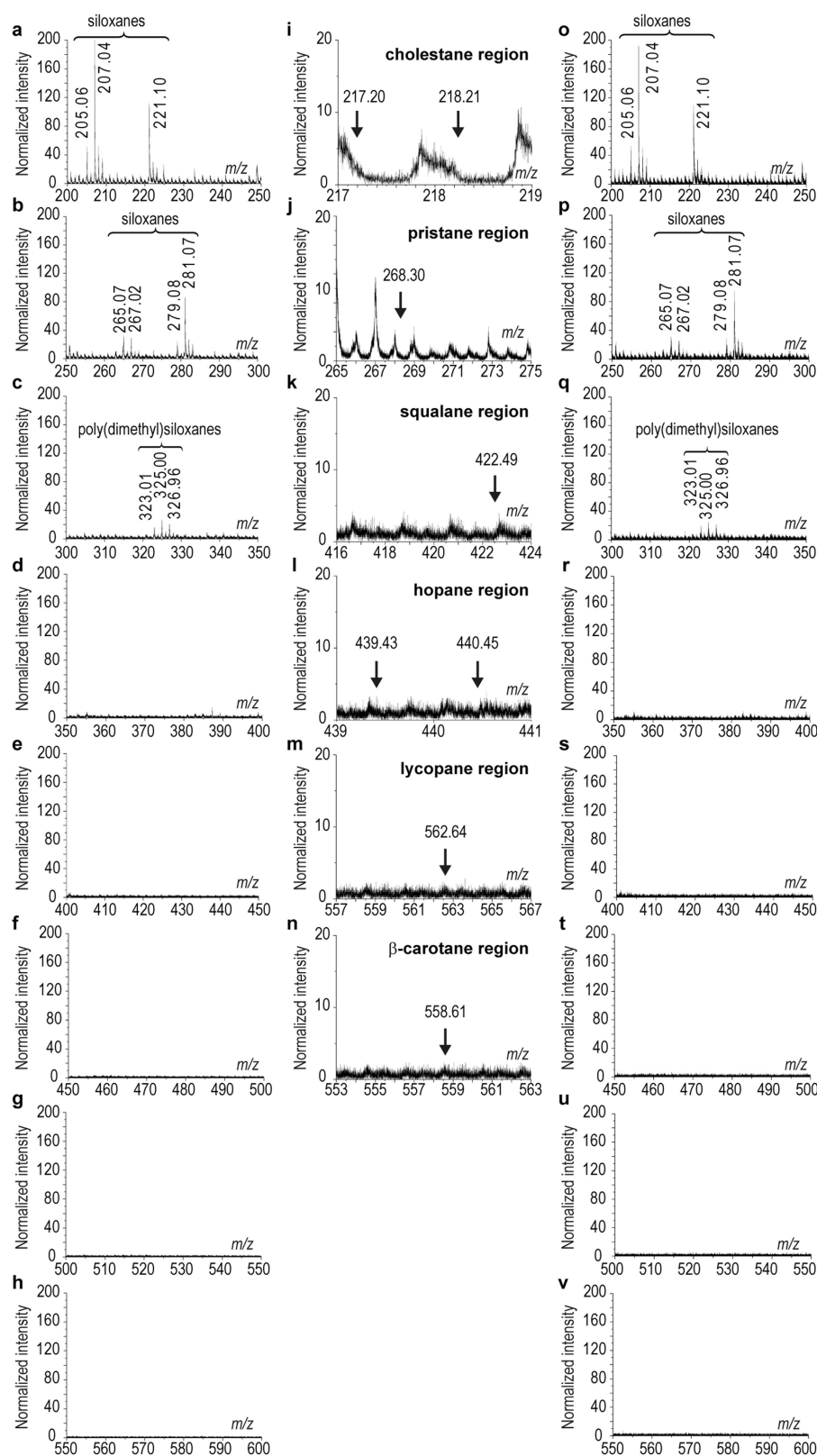
Extended Data Fig. 6 | Systematic association of Fe-rich saponite with tryptophan. **a**, Optical view of an olivine kernel in the mesh serpentine, as shown in Extended Data Fig. 1b. **b**, **c**, SEM-BSE images collected at 15 kV in the area indicated by the green box in **a**. The image in panel **b** was modified from ref. ¹⁵, Springer Nature Limited. The orange dashed line and associated arrow in **b**, respectively, provide the location and front face of the FIB foil (Extended Data Fig. 7) milled for the TEM observations displayed in **d**. **c**, Magnified view of the area represented by the green box in **b**. Textures and chemical contrasts allow the recognition of each mineral phase previously characterized by S-FTIR and electron microprobe analysis, namely magnetite, olivine, mesh serpentine, Fe-rich serpentine and Fe-rich saponite. **d**, TEM image that shows the Fe-rich saponite layers appear perpendicular to the interface between the olivine kernel and the

mesh serpentine, with lizardite crystallites appearing in black. The Fe-rich saponite lamellae are mainly subparallel although some sheet distortions are visible. **e**, Associated S-DUV full-field fluorescence images collected after excitation (λ_{exc}) at 275 nm in the range 327–353, 370–410 and 412–438 nm. For all of these wavelengths, the UV autofluorescence emission is shown at the interface between the olivine crystal and serpentine where the Fe-rich saponite is located. **f**, TOF-SIMS ion image collected in the $500 \times 500 \mu\text{m}^2$ area displayed in **a**, and showing the distribution of the characteristic fragment ions of tryptophan²² (Supplementary Table 2), hence confirming its presence close to the olivine kernel. Evidence for the absence of common biomarkers^{23–25} can be found in the detailed spectra provided in Extended Data Fig. 8o–v. The green arrows provide sample orientation.



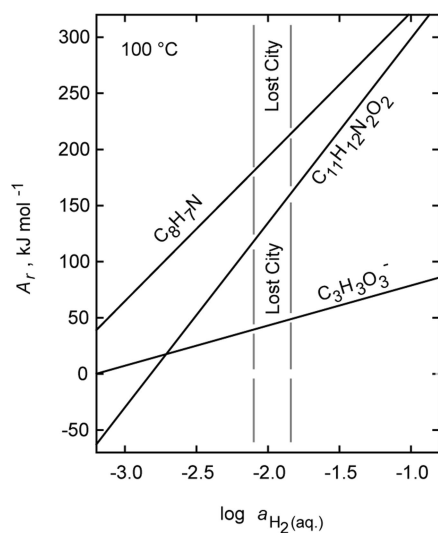
Extended Data Fig. 7 | SEM image sequence that illustrates FIB milling, which allowed cross-sectional visualization of the interfaces between the UV-fluorescent Fe-rich saponite, the olivine kernel and the mesh serpentine. **a**, SEM-SESI view at low magnification of the UV-fluorescent area shown in Extended Data Fig. 6. The white arrow denotes the region of interest, where an ultrathin foil was milled for TEM observations (Extended Data Fig. 6d). The green arrow indicates the orientation of the

sample. **b**, SEM-SESI view of the region of interest coated with a carbon-protective layer. The orange arrow designates the milling direction. **c**, SEM-SESI image showing the FIB milled trench at the leading edge of the region of interest. **d**, SEM-BSE image showing the front face of the milled section. Similarly to Extended Data Fig. 4, it revealed marked textural contrasts between the mineral phases, the Fe-rich saponite presenting the highest porosity.

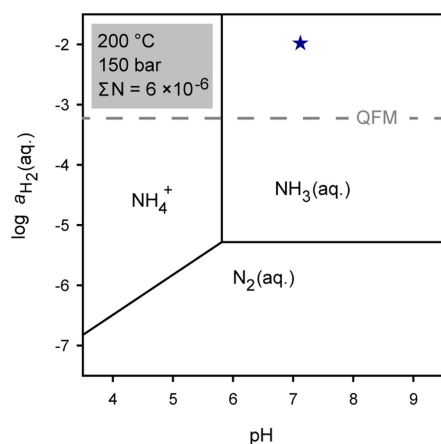


Extended Data Fig. 8 | Enlarged views of the positive TOF-SIMS spectra collected in the Fe-rich saponite. a–n, Enlarged views of the positive TOF-SIMS spectrum from Fig. 2d reconstructed from the region displaying the highest count rate in Fig. 2b. **i–n,** Selected magnified views of this TOF-SIMS spectrum showing regions in which the peaks of fragment ions characteristic of isoprenoids such as pristane ($C_{19}H_{40}$), squalane ($C_{30}H_{50}$) and lycopane ($C_{40}H_{56}$) along with polycyclic compounds (cholestane, $C_{27}H_{48}$; β -carotane, $C_{40}H_{56}$ and hopanoids) should be found if present^{23–25}. These biomarkers were previously detected

using gas chromatography in the bulk rock¹³ but are not detected locally in our mineral assemblage. **o–v,** Enlarged views of the positive TOF-SIMS spectra reconstructed from the region displaying the highest count rate in Extended Data Fig. 6f. **a–c** and **o–q** exhibit peaks of fragments ions characteristic of siloxane, a common plasticizer contaminant. No notable peaks can be found in the 350–450 m/z regions (**d, e, r** and **s**), in which the aliphatic fraction (including fragment ions from the sterane and hopane classes and from alkanes and monocyclic alkanes) is expressed in positive TOF-SIMS spectra^{23–25}.



Extended Data Fig. 9 | Chemical affinity as a function of the logarithm of the activity of aqueous dihydrogen ($\text{H}_2(\text{aq.})$) for the reactions corresponding to the abiotic synthesis of pyruvate ($\text{C}_3\text{H}_3\text{O}_3^-$), indole ($\text{C}_8\text{H}_7\text{N}$), and tryptophan ($\text{C}_{11}\text{H}_{12}\text{N}_2\text{O}_2$). The vertical lines indicate the range of H_2 activities reported for the moderate-temperature fluids of the Lost City hydrothermal field³⁸. See Supplementary Information.



Extended Data Fig. 10 | Activity diagram of aqueous dihydrogen $\text{H}_2(\text{aq.})$ depicting, as a function of pH, the fields of relative predominance of nitrogen species at 200 °C and 150 bar and considering a total nitrogen amount (ΣN) of $6 \times 10^{-6} \text{ M}$. The limits between two predominance fields have been drawn for equal activities of the nitrogen species (that is, ammonium (NH_4^+), aqueous ammonia ($\text{NH}_3(\text{aq.})$) and aqueous dinitrogen ($\text{N}_2(\text{aq.})$)). The blue star indicates conditions encountered at depth in the Atlantis Massif by considering a mean H_2 activity of 10.5 mM ³⁸ and a pH of 7.12, calculated with tremolite-chrysotile-diopside as the alteration assemblage consistent with hydrothermal fluid composition³⁷. The activity of water was assumed equal to 1. Diagram shows that $\text{NH}_3(\text{aq.})$ is thermodynamically favoured at 200 °C. QFM, quartz-fayalite-magnetite mineral buffer.

Amphioxus functional genomics and the origins of vertebrate gene regulation

Ferdinand Marlétaz^{1,2,41}, Panos N. Firas^{3,41}, Ignacio Maeso^{3,41*}, Juan J. Tena^{3,41}, Ozren Bogdanovic^{4,5,6,41}, Malcolm Perry^{7,8,41}, Christopher D. R. Wyatt^{9,10}, Elisa de la Calle-Mustienes³, Stephanie Bertrand¹¹, Demian Burguera^{9,12}, Rafael D. Acemel³, Simon J. van Heeringen¹³, Silvia Naranjo³, Carlos Herrera-Ubeda¹², Ksenia Skvortsova⁴, Sandra Jimenez-Gancedo³, Daniel Aldea¹¹, Yamile Marquez⁹, Lorena Buono³, Iryna Kozmikova¹⁴, Jon Permanyer⁹, Alexandra Louis^{15,16,17}, Beatriz Albuixech-Crespo¹², Yann Le Petillon¹¹, Anthony Leon¹¹, Lucie Subirana¹¹, Piotr J. Balwierz^{7,8}, Paul Edward Duckett⁴, Ensieh Farahani³, Jean-Marc Aury¹⁸, Sophie Mangelot¹⁸, Patrick Wincker¹⁹, Ricard Albalat²⁰, Èlia Benito-Gutiérrez²¹, Cristian Cañestro²⁰, Filipe Castro²², Salvatore D'Aniello²³, David E. K. Ferrier²⁴, Shengfeng Huang²⁵, Vincent Laudet¹¹, Gabriel A. B. Marais²⁶, Pierre Pontarotti²⁷, Michael Schubert²⁸, Hervé Seitz²⁹, Ildiko Somorjai³⁰, Tokiharu Takahashi³¹, Olivier Mirabeau³², Anlong Xu^{25,33}, Jr-Kai Yu³⁴, Piero Carninci^{35,36}, Juan Ramon Martinez-Morales³, Hugues Roest Crollius^{15,16,17}, Zbynek Kozmik¹⁴, Matthew T. Weirauch^{37,38}, Jordi Garcia-Fernández¹², Ryan Lister^{6,39}, Boris Lenhard^{7,8,40}, Peter W. H. Holland¹, Hector Escriva^{11*}, Jose Luis Gómez-Skarmeta^{3*} & Manuel Irimia^{9,10*}

Vertebrates have greatly elaborated the basic chordate body plan and evolved highly distinctive genomes that have been sculpted by two whole-genome duplications. Here we sequence the genome of the Mediterranean amphioxus (*Branchiostoma lanceolatum*) and characterize DNA methylation, chromatin accessibility, histone modifications and transcripts across multiple developmental stages and adult tissues to investigate the evolution of the regulation of the chordate genome. Comparisons with vertebrates identify an intermediate stage in the evolution of differentially methylated enhancers, and a high conservation of gene expression and its *cis*-regulatory logic between amphioxus and vertebrates that occurs maximally at an earlier mid-embryonic phylotypic period. We analyse regulatory evolution after whole-genome duplications, and find that—in vertebrates—over 80% of broadly expressed gene families with multiple paralogues derived from whole-genome duplications have members that restricted their ancestral expression, and underwent specialization rather than subfunctionalization. Counter-intuitively, paralogues that restricted their expression increased the complexity of their regulatory landscapes. These data pave the way for a better understanding of the regulatory principles that underlie key vertebrate innovations.

All vertebrates share multiple morphological and genomic novelties¹. The most prominent genomic difference between vertebrates and non-vertebrate chordates is the reshaping of the gene complement that followed the two rounds of whole genome duplication (WGD)—the 2R hypothesis—that occurred at the base of the vertebrate lineage^{2,3}. These large-scale mutational events are hypothesized to have

facilitated the evolution of vertebrate morphological innovations, at least in part through the preferential retention of 'developmental' gene families and transcription factors after duplication^{3,4}. However, duplicate genes and their associated regulatory elements were initially identical and could not drive innovation without regulatory and/or protein-coding changes.

¹Department of Zoology, University of Oxford, Oxford, UK. ²Molecular Genetics Unit, Okinawa Institute of Science and Technology Graduate University, Onna-son, Japan. ³Centro Andaluz de Biología del Desarrollo (CABD), CSIC-Universidad Pablo de Olavide-Junta de Andalucía, Seville, Spain. ⁴Genomics and Epigenetics Division, Garvan Institute of Medical Research, Sydney, New South Wales, Australia. ⁵St Vincent's Clinical School, Faculty of Medicine, University of New South Wales, Sydney, New South Wales, Australia. ⁶Australian Research Council Centre of Excellence in Plant Energy Biology, School of Molecular Sciences, The University of Western Australia, Crawley, Western Australia, Australia. ⁷Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London, UK. ⁸Computational Regulatory Genomics, MRC London Institute of Medical Sciences, London, UK. ⁹Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. ¹⁰Universitat Pompeu Fabra (UPF), Barcelona, Spain. ¹¹Biologie Intégrative des Organismes Marins, BIOM, Observatoire Océanologique, CNRS and Sorbonne Université, Banyuls sur Mer, France. ¹²Department of Genetics, Microbiology and Statistics, Faculty of Biology, and Institut de Biomedicina (IBUB), University of Barcelona, Barcelona, Spain. ¹³Department of Molecular Developmental Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences, Radboud University, Nijmegen, The Netherlands. ¹⁴Institute of Molecular Genetics of the Czech Academy of Sciences, Prague, Czech Republic. ¹⁵Institut de Biologie de l'ENS, IBENS, Ecole Normale Supérieure, Paris, France. ¹⁶Inserm, U1024, Paris, France. ¹⁷CNRS, UMR 8197, Paris, France. ¹⁸Genoscope, Institut de biologie François-Jacob, Commissariat à l'Energie Atomique (CEA), Université Paris-Saclay, Evry, France. ¹⁹Génomique Métabolique, Genoscope, Institut de biologie François Jacob, Commissariat à l'Energie Atomique (CEA), CNRS, Université Evry, Université Paris-Saclay, Evry, France. ²⁰Department of Genetics, Microbiology and Statistics, Faculty of Biology and Institut de Recerca de la Biodiversitat (IRBio), University of Barcelona, Barcelona, Spain. ²¹Department of Zoology, University of Cambridge, Cambridge, UK. ²²Interdisciplinary Centre of Marine and Environmental Research (CIIMAR/CIMAR) and Faculty of Sciences (FCUP), Department of Biology, University of Porto, Porto, Portugal. ²³Biology and Evolution of Marine Organisms, Stazione Zoologica Anton Dohrn Napoli, Naples, Italy. ²⁴The Scottish Oceans Institute, Gatty Marine Laboratory, University of St Andrews, St Andrews, UK. ²⁵State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China. ²⁶Laboratoire de Biométrie et Biologie Evolutive (UMR 5558), CNRS and Université Lyon 1, Villeurbanne, France. ²⁷IRD, APHM, Microbe, Evolution, Phylogénie, Infection, IHU Méditerranée Infection and CNRS, Aix Marseille University, Marseille, France. ²⁸Sorbonne Université, CNRS, Laboratoire de Biologie du Développement de Villefranche-sur-Mer, Institut de la Mer de Villefranche-sur-Mer, Villefranche-sur-Mer, France. ²⁹UMR 9002 CNRS, Institut de Génétique Humaine, Université de Montpellier, Montpellier, France. ³⁰Biomedical Sciences Research Complex, School of Biology, University of St Andrews, St Andrews, UK. ³¹School of Medical Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK. ³²INSERM U830, Équipe Labellisée LNCC, SIREDO Oncology Centre, Institut Curie, PSL Research University, Paris, France. ³³School of Life Sciences, Beijing University of Chinese Medicine, Beijing, China. ³⁴Institute of Cellular and Organismic Biology, Academia Sinica, Taipei, Taiwan. ³⁵RIKEN Center for Life Science Technologies (Division of Genomic Technologies) (CLST DGT), Yokohama, Japan. ³⁶Laboratory for Transcriptome Technology, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. ³⁷Center for Autoimmune Genomics and Etiology, Divisions of Biomedical Informatics and Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. ³⁸Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA. ³⁹Harry Perkins Institute of Medical Research, Nedlands, Western Australia, Australia. ⁴⁰Sars International Centre for Marine Molecular Biology, University of Bergen, Bergen, Norway. ⁴¹These authors contributed equally: Ferdinand Marlétaz, Panos N. Firas, Ignacio Maeso, Juan J. Tena, Ozren Bogdanovic, Malcolm Perry. *e-mail: nacho.maeso@gmail.com; hescriva@obs-banyuls.fr; jlgomska@upo.es; mirimia@gmail.com

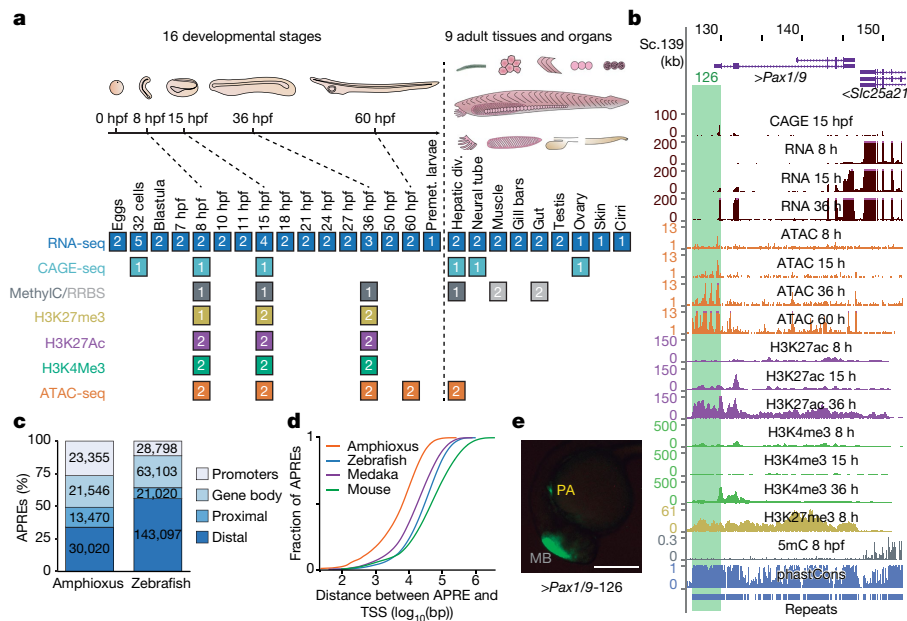


Fig. 1 | Functional genome annotation of amphioxus. **a**, Summary of the 94 amphioxus samples generated in this study, comprising eight functional-genomic datasets. The number of biological replicates is indicated for each sample type. div., diverticulum; Methylation/RRBS, methylC sequencing and reduced representation bisulfite sequencing; Premet., premetamorphic. **b**, Genome browser excerpt showing a selection of available tracks, including gene annotation, sequence conservation (using phastCons), repeats and several epigenomic and transcriptomic datasets. Green rectangle highlights the APRE tested in **e**. **c**, Numbers and proportions of amphioxus and zebrafish APREs according to their

genomic location. Promoters, within 1-kbp upstream and 0.5-kbp downstream of an annotated TSS; gene body, within an orthology-supported gene; proximal, within 5-kbp upstream of (but not overlapping with) a TSS; distal, not in the aforementioned categories. **d**, Cumulative distributions of the distance between each APRE and the closest annotated TSS in each species. **e**, Lateral view of a representative transgenic zebrafish 26-hpf embryo showing GFP expression driven by an amphioxus APRE associated with *Pax1/9* (*Pax1/9-126*, highlighted in **b**) in pharyngeal arches (PA; $n = 4/4$). Positive-control enhancer was expressed in the midbrain (MB). Scale bar, 250 μ m.

To date, the effect of vertebrate WGDs on gene regulation have remained poorly understood—both in terms of the fates of duplicate genes and the acquisition of the unique genomic traits that are characteristic of vertebrates. These traits include numerous features that are often associated with gene regulation, such as unusually large intergenic and intronic regions^{5,6}, high global 5-methylcytosine (5mC) content and 5mC-dependent regulation of embryonic transcriptional enhancers⁷. To investigate these traits, appropriate species must be used for comparisons. Previous studies have largely focused on phylogenetic distances that are either too short (such as human versus mouse) or too long (such as human versus fly or nematode), resulting in limited insights. In the first case, comparisons among closely related species (for example, between mammals^{8–11})—for which the orthology of non-coding regions can be readily determined from genomic alignments—have allowed fine-grained analyses of the evolution of transcription-factor binding. In the second case, three-way comparisons of human, fly and nematode by the modENCODE consortium revealed no detectable conservation at the *cis*-regulatory level¹² and very little conservation of gene expression¹³. Moreover, the genomes of flies and nematodes are highly derived^{14–16}. Thus, we lack comprehensive functional genomic data from a slow-evolving, closely related outgroup that would enable an in-depth investigation of the origins of the vertebrate regulatory genome and of the effect of WGDs on gene regulation.

Unlike flies, nematodes and most non-vertebrates, amphioxus belongs to the chordate phylum. Therefore, although amphioxus lacks the specializations and innovations of vertebrates, it shares with them a basic body plan and has multiple organs and structures homologous to those of vertebrates¹. For these reasons, amphioxus has widely been used as a reference outgroup to infer ancestral versus novel features during vertebrate evolution. Here, we undertook a comprehensive study of the transcriptome and regulatory genome of amphioxus to investigate how the unique functional genome architecture of vertebrates evolved.

Functional genome annotation of amphioxus

We generated an exhaustive resource of genomic, epigenomic and transcriptomic data for the Mediterranean amphioxus (*B. lanceolatum*), comprising a total of 52 sample types (Fig. 1a and Supplementary Data 2, datasets 1–5). These datasets were mapped to a *B. lanceolatum* genome that was sequenced and assembled de novo, with 150 \times coverage, a total size of 495.4 Mbp, a scaffold N50 of 1.29 Mbp and 4% gaps (Extended Data Fig. 1a–c). To facilitate access by the research community, we integrated these resources into a UCSC Genome Browser track hub (Fig. 1b; available at <http://amphiencode.github.io/Data/>), together with an intra-cephalochordate sequence conservation track and a comprehensive annotation of repetitive elements (Extended Data Fig. 1d–f) and long non-coding RNAs (Extended Data Fig. 1g and Supplementary Data 2, dataset 6). To enable broader evolutionary comparisons, we reconstructed orthologous gene families for multiple vertebrate and non-vertebrate species (Supplementary Data 2, dataset 7), generated several equivalent datasets for zebrafish and medaka (Extended Data Fig. 2a), and built a dedicated server for synteny comparisons (Extended Data Fig. 1h).

A comprehensive functional annotation of the *B. lanceolatum* genome identified 88,391 putative *cis*-regulatory elements of DNA as defined by assay for transposase-accessible chromatin using sequencing (ATAC-seq) (these elements are hereafter referred to as APREs), as well as 20,569 protein-coding genes supported by orthology. We divided the APREs into promoters—around transcription start sites (TSSs), which were highly supported by cap analysis gene-expression sequencing (CAGE-seq) data, Extended Data Fig. 2b—and gene-body, proximal and distal APREs (Fig. 1c). Equivalent analyses using zebrafish data yielded 256,018 potential regulatory regions, with a significantly higher proportion of these being distal APREs (Fig. 1c; $P < 2.2 \times 10^{-16}$, one-sided Fisher's exact test). A significantly larger global TSS distance in APREs was observed for all vertebrates compared to amphioxus (Fig. 1d), even after correcting for differences in average intergenic

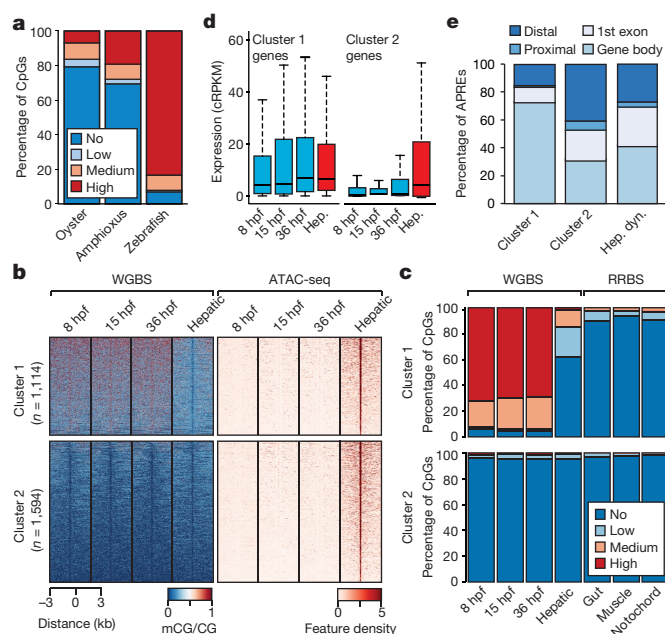


Fig. 2 | 5mC patterns and dynamics in the amphioxus genome. **a**, Percentage of methylated CpG dinucleotides in oyster (mantle, $n = 14,779,123$), amphioxus (8 hpf, $n = 19,657,388$) and zebrafish (1,000-cell stage, $n = 38,989,847$) samples. Low, >0 –20%; medium, 20–80%; high, $>80\%$. **b**, k -means clustering ($n = 2$) of 5mC signal over hepatic-specific APREs. **c**, Percentage of methylated CpG dinucleotides as assessed by whole-genome bisulfite sequencing (WGBS) and reduced representation bisulfite sequencing (RRBS) in embryos and adult tissues in APREs from **b**. **d**, Distribution of expression levels for genes associated with APREs displaying distinct 5mC patterns in **b**. Cluster 1: 1,114 genes; cluster 2: 1,594 genes. cRPKM, corrected (per mappability) reads per kb of mappable positions and million reads. Hep, hepatic diverticulum. **e**, Genomic distribution of regions with distinct 5mC patterns from **b**. Hep. dyn., dynamic APREs active in the hepatic diverticulum.

length among species (Extended Data Fig. 2c; $P < 2.2 \times 10^{-16}$ for all vertebrate-versus-amphioxus comparisons, one-sided Mann–Whitney tests). Amphioxus APREs showed enrichment for enhancer-associated

chromatin marks (Extended Data Fig. 2d), which were highly dynamic during embryo development (Extended Data Fig. 2e–g), and consistently drove GFP expression in zebrafish or amphioxus transgenic assays (93% (14/15), Fig. 1e and Extended Data Fig. 2h, i). Moreover, 89% (32/36) of previously reported amphioxus enhancers overlapped APREs defined by our data. Therefore, a large fraction of APREs probably act as developmentally regulated transcriptional enhancers.

Disentangling vertebrate bidirectional promoters

Analyses of core promoters, defined by CAGE-seq, at single-nucleotide resolution revealed that amphioxus promoters display a mixture of pan-metazoan, pan-vertebrate and unique features (Extended Data Fig. 3 and Supplementary Information). These analyses also identified that 25% (3,950/15,884) of neighbouring protein-coding genes were arranged in bidirectional promoters. Bidirectional promoters were most common among ubiquitous promoters (Extended Data Fig. 4a), displayed a marked periodicity in the distance between promoters (Extended Data Fig. 4b, c) and were associated with genes that were significantly enriched in housekeeping functions (Extended Data Fig. 4d). Notably, the fraction of bidirectional promoters defined by CAGE-seq decreased progressively from amphioxus to mouse (12.83% (1,752/13,654)) and to zebrafish (7.84% (1,098/14,014)), which suggests a disentanglement of ancestral bidirectional promoters after each round of WGD (two in tetrapods and three in teleosts). Consistently, the majority of a set of 372 putatively ancestral, bidirectional promoters were lost in vertebrates—particularly in stem vertebrates (54.5%)—with only very few amphioxus-specific losses (5.3%) (Extended Data Fig. 4e, f).

Developmental DNA demethylation of APREs

Similar to other non-vertebrates^{17–19}, the amphioxus genome exhibited very low levels of CpG methylation (Fig. 2a); nearly all of the 5mC occurred in gene bodies, in which the proportion of methylated CpGs correlated positively with gene-expression levels but negatively with the density of H3K27me3 and H3K4me3 histone marks and CpG dinucleotides (Extended Data Fig. 5a–c). However, as in zebrafish and frogs⁷, global levels of 5mC displayed a decrease during development (Extended Data Fig. 5d–g), coinciding with the onset of expression of the amphioxus orthologue of TET demethylase (Extended Data Fig. 5h).

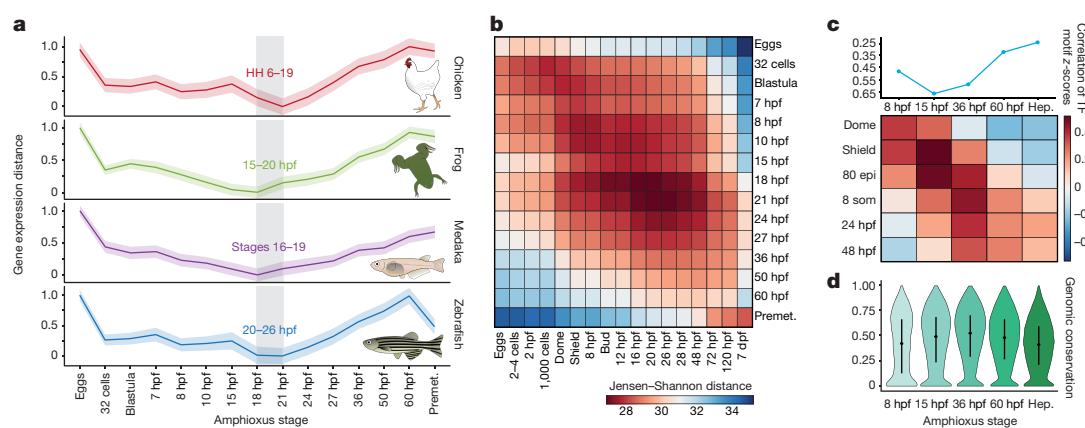


Fig. 3 | The hourglass model and chordate embryogenesis. **a**, Stages of minimal transcriptomic divergence (using the Jensen–Shannon distance metric) from four vertebrate species to each amphioxus stage. The grey box outlines the period of minimal divergence, with the corresponding vertebrate periods indicated (the range is given by the two less divergent stages). Dispersions correspond to the standard deviation computed on 100 bootstrap re-samplings of the orthologue sets (amphioxus–chicken: 5,720; amphioxus–zebrafish: 5,673; amphioxus–frog: 5,883; and amphioxus–medaka: 5,288). HH, Hamburger–Hamilton stage. **b**, Heat map of pairwise transcriptomic Jensen–Shannon distances between amphioxus (vertical) and zebrafish (horizontal) stages. A smaller distance (red) indicates higher similarity. dpf, days post-fertilization. **c**, Zebrafish and amphioxus pairwise Pearson correlation of relative enrichment z-scores for transcription-factor (TF) motifs in dynamic APREs, active at different developmental stages. Top, maximal correlation for each amphioxus stage against the zebrafish stages. Bottom, heat map with all pairwise correlations. 80 epi, 80% epiboly stage; 8 som, 8-somite stage. **d**, Sequence conservation levels within the cephalochordates of active APREs at each developmental stage, visualized as the distribution of average phastCons scores. The number of APREs at 8 hpf = 5,282; at 15 hpf = 17,387; at 36 hpf = 21,089; at 60 hpf = 22,674; and in hepatic diverticulum (hep) = 16,551. Dots correspond to the mean values and lines represent the interquartile range.

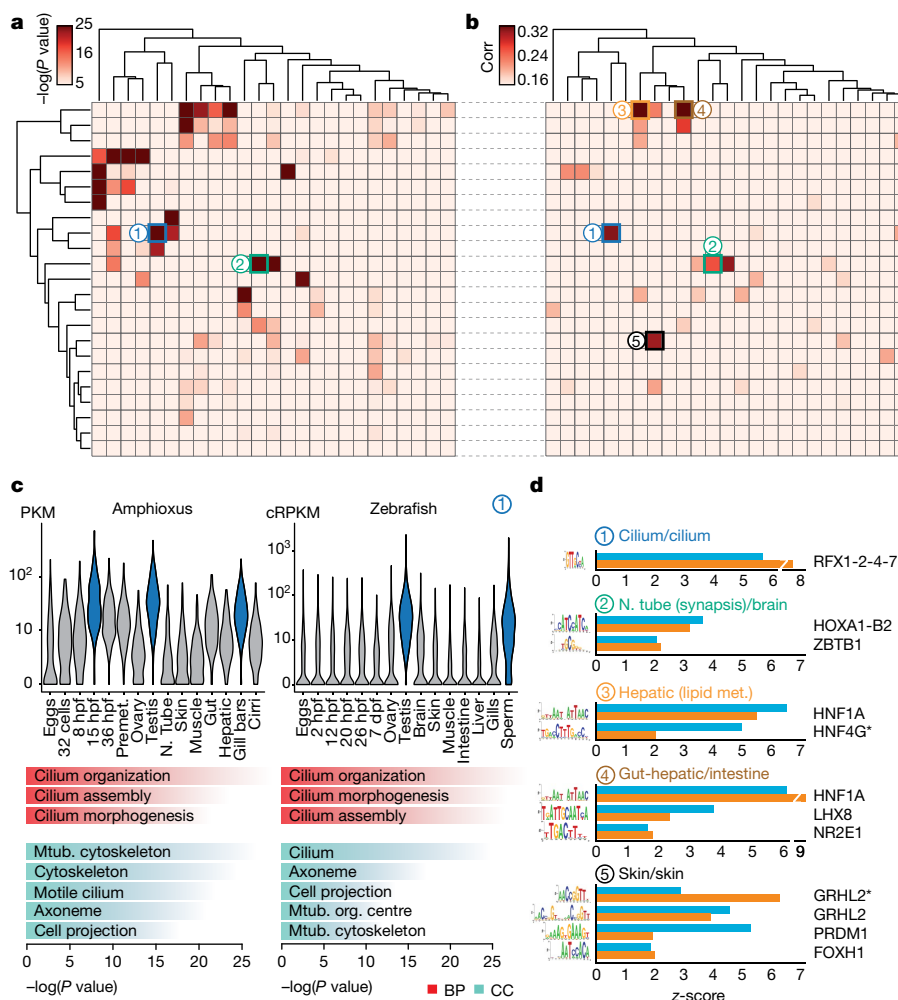


Fig. 4 | Transcriptomic and *cis*-regulatory conservation of adult chordate tissues.
a, Heat map showing the level of raw statistical significance of orthologous gene overlap between modules produced by weighted correlation network analysis (WGCNA), from amphioxus (vertical) and zebrafish (horizontal) as derived from upper-tail hypergeometric tests. **b**, Heat map of all pairwise Pearson correlations (corr) between the modules of the two species, based on the relative z-scores of transcription-factor motifs for each module (242 super-families of motifs). Modules are clustered as in **a**. **c**, Distribution of expression values (cRPKMs) for all genes within the cilium modules across each sample (top), and enriched Gene Ontology terms within each module (bottom) for a pair of modules (labelled '1' in **b**; 1,681 and 261 genes in zebrafish and amphioxus, respectively). BP, biological process; CC, cellular component. *P* values correspond to uncorrected two-sided Fisher's exact tests as provided by topGO. Mtub., microtubule; N. tube, neural tube; org., organizing. **d**, Transcription-factor binding-site motifs with high z-scores from highly correlated pairs of modules between zebrafish (blue) and amphioxus (orange). Numbers correspond to those circles in **b**. RFX1-2-4-7 denotes RFX1, RFX2, RFX4 and RFX7; HOXA1-B2 denotes HOXA1 and HOXB2; asterisk denotes an alternative motif.

To assess whether these 5mC dynamics may have regulatory potential, we identified adult hepatic diverticulum-specific APREs that are inactive during development. Unlike embryo-specific APREs (Extended Data Fig. 6a), the clustering of these adult APREs on the basis of 5mC content revealed two distinct subsets, one with hepatic-specific and one with constitutive hypomethylation (Fig. 2b). Differentially methylated APREs (cluster 1) also displayed robust hypomethylation in other adult tissues (Fig. 2c), which suggests that demethylation at these APREs occurs organism-wide. Both groups of hepatic-specific APREs were enriched for binding sites of liver-specific transcription factors—such as Hnf4a—as well as broadly expressed transcription factors such as Foxa (Extended Data Fig. 6b), which is a pioneer factor that participates in 5mC removal at regulatory regions in mammals²⁰.

APREs from both clusters were preferentially associated with genes with metabolic functions (Extended Data Fig. 6c). However, only APREs with hepatic-specific hypomethylation (cluster 1) were primarily associated with genes that displayed steady widespread expression (Fig. 2d and Extended Data Fig. 6d, e); these APREs were mainly located within gene bodies (Fig. 2e). These data suggest that demethylation of these APREs may contribute to their identification as adult-specific, transcriptional *cis*-regulatory elements within continuously hypermethylated gene-body contexts, which is characteristic of non-vertebrate species. Fourteen zebrafish gene families contained differentially methylated APREs in introns that are orthologous to those identified in amphioxus—amongst these are four genes that encode components of the Hippo pathway, including the transcriptional effectors Yap (*yap1* and *wvtr1*) and Tead (*tead1a* and *tead3a*) (Extended Data Fig. 6f, g).

The hourglass model and chordate embryogenesis

Previous comparative analyses among vertebrate transcriptomes^{21,22} showed a developmental period of maximal similarity in gene expression that coincides with the so-called phylotypic period, consistent with the hourglass model²³. However, similar comparisons with tunicates and amphioxus have thus far not resolved a phylotypic period shared across all chordates²². Pairwise comparisons of stage-specific RNA sequencing (RNA-seq) data from developmental time courses of amphioxus against zebrafish, medaka, frog (*Xenopus tropicalis*) and chicken revealed a consistent period of highest similarity (Fig. 3a, b and Extended Data Fig. 7) that occurred slightly earlier than those reported for vertebrates; in amphioxus, this corresponds to the neurula at the 4–7-somite stage (18–21 hours post fertilization (hpf)). At the regulatory level, pairwise comparisons between the relative enrichment of transcription-factor motifs in sets of dynamic APREs that were active at each stage were also consistent with an earlier hourglass model²⁴ (Fig. 3c). By contrast, at a shorter timescale, comparisons between different species of amphioxus showed that the sequence conservation for the same APREs was higher after the putative chordate phylotypic period (Fig. 3d).

Regulatory conservation shapes chordate body plan

Additional comparisons of embryo transcriptomes and neighbourhood analysis of conserved co-expression²⁵ showed a high conservation of developmental and global expression patterns and of gene functions between amphioxus and vertebrates (Extended Data Fig. 8 and Supplementary Information). Further pairwise comparison of co-regulated gene modules across tissues between amphioxus and zebrafish revealed multiple pairs with highly significant levels of orthologue overlap (Fig. 4a). These included modules with conserved tissue-specific

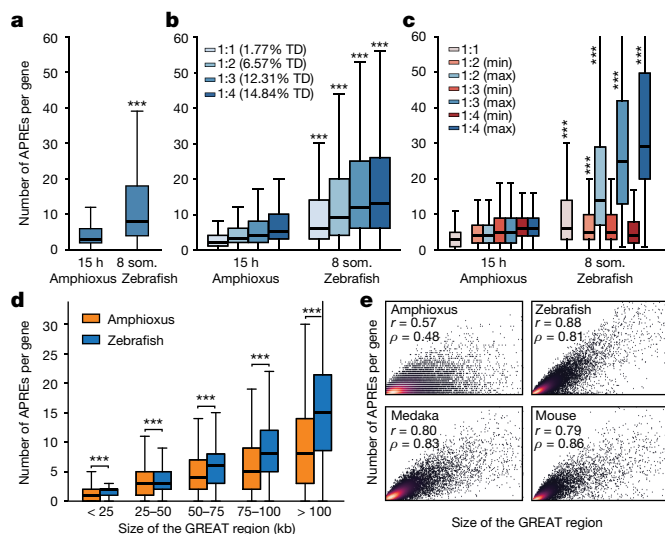


Fig. 5 | Higher regulatory complexity in vertebrate regulatory landscapes. **a**, Distribution of the number of APREs within the regulatory landscape of each gene (as estimated by GREAT), at comparable pre-phylogenetic developmental stages (15 hpf for amphioxus and 8 somites for zebrafish). $n = 6,047$ and $9,239$ genes for amphioxus and zebrafish, respectively. **b**, As in **a**, but with gene families separated according to the number of retained ohnologues per family in vertebrates (from 1 to 4, using mouse as a reference). The percentage of developmental regulatory genes (TD, trans-dev) in each category is indicated. **c**, As in **b**, but only the genes with the lowest ('min', in red) and the highest ('max', in blue) number of APREs are plotted for each ohnologue family. **d**, Distributions of the number of APREs per gene among subsets of amphioxus and zebrafish genes matched by GREAT-region size (± 500 bp) and binned by size as indicated. **e**, Density scatter plot of the number of APREs (y axis) versus the size of the GREAT region (x axis) per gene and species. Pearson (r) and Spearman (ρ) correlation coefficients are indicated. Sample sizes: amphioxus, 20,053; zebrafish, 20,569; medaka, 15,978; mouse, 18,838. **a–d**, *** $P < 0.001$; one-sided Mann–Whitney tests of the zebrafish distribution versus the equivalent amphioxus distribution. Exact P values and sample sizes are provided in Supplementary Data 2, dataset 8.

expression that were enriched for coherent Gene Ontology categories, including genes with high expression in organs with ciliated cells (for example, spermatozoa and gill bars) (labelled '1' in Fig. 4a–c) as well as neural, muscle, gut, liver, skin and metabolism-related modules (Supplementary Data 1). We also found a significant positive correlation between relative motif-enrichment scores for many pairs of modules (Fig. 4b); the most-enriched transcription-factor motifs within each cluster were highly consistent between amphioxus and zebrafish (Fig. 4d).

Higher regulatory information in vertebrate genomes

To investigate the effect of WGDs on the evolution of vertebrate gene regulation, we first asked whether the number of putative regulatory regions per gene is higher in vertebrates than in amphioxus. We observed significantly more APREs in the regulatory landscape of each gene (as defined by the 'Genomic Regions Enrichment of Annotations Tool' (GREAT)²⁶) in zebrafish than in amphioxus (Fig. 5a). This difference is particularly evident for gene families that have retained multiple copies after WGD (known as ohnologues; Fig. 5b), for which the number of APREs is very uneven between copies, with marked regulatory expansions observed for some ohnologues (Fig. 5c). The same patterns were detected for all developmental stages of amphioxus and zebrafish, as well as for medaka and mouse genomes, and were highly robust to down-sampling of ATAC-seq coverage in vertebrates (Extended Data Fig. 9a–c). We also detected a higher number of peaks associated with regulatory genes ('trans-dev' genes that are involved in the regulation of embryonic development) compared to housekeeping genes in all species (Extended Data Fig. 9d), consistent

with the higher frequency of retention of trans-dev genes in multiple copies after WGD³ (Fig. 5b). Comparison of regulatory landscapes—determined experimentally using circular chromosome conformation capture followed by sequencing (4C-seq)—for 58 genes from 11 trans-dev gene families in amphioxus, zebrafish and mouse showed similar results (Extended Data Fig. 9e).

As expected, the higher number of APREs in zebrafish was associated with larger intergenic regions in this species (Extended Data Fig. 9f). However, the differences in APRE complements were not attributable only to an increase in genome size in vertebrates, as subsets of amphioxus and zebrafish genes with matched distributions of GREAT or intergenic-region lengths also displayed a higher number of APREs in zebrafish (Extended Data Fig. 9g, h). Further investigation of matched distributions showed that these differences were particularly great in genes with large regulatory landscapes (>50 kb) (Fig. 5d). Thus, larger regions in amphioxus did not scale at the same rate as in vertebrates in terms of regulatory complexity (Fig. 5e), which is consistent with the overall lower proportion of distal APREs identified in this species (Fig. 1c, d). In summary, these analyses reveal a large increase in the number of regulatory regions during vertebrate evolution (and/or a decrease in these regions in amphioxus)—particularly of distal regulatory elements—and that this trend is enhanced for specific gene copies retained after the WGDs, pointing to unequal rates of regulatory evolution for different ohnologues.

More-complex regulation in specialized ohnologues

The duplication–degeneration–complementation (DDC) model hypothesizes that the retention of duplicate genes could be driven by reciprocal loss of regulatory elements and restriction of paralogues to distinct subsets of the ancestral expression pattern²⁷. In particular, the DDC model predicts that individual paralogues would each have more restricted expression than an unduplicated outgroup, but that their summation would not. To test this, we binarized the expression ('on' or 'off') of each gene in nine homologous expression domains in amphioxus, zebrafish, frog and mouse (Fig. 6a). When comparing genes that returned to single-copy status after WGDs, we detected no expression bias between amphioxus and vertebrates (Fig. 6a, b and Extended Data Fig. 10a, b). By contrast, when vertebrate ohnologues were compared to their single amphioxus orthologues, the distributions were strongly skewed and many vertebrate genes displayed far more restricted expression domains (Fig. 6b and Extended Data Fig. 10a, b; similar results were obtained by comparing τ values²⁸, Extended Data Fig. 10c–e). The symmetrical pattern was fully recovered when the expression of all vertebrate members was combined, or when the raw expression values were summed for each member within a paralogy group (Fig. 6a, b and Extended Data Fig. 10a, b).

Although the above findings are consistent with the DDC model, they are also compatible with an alternative model in which a subset of duplicate genes becomes more 'specialized' in expression pattern while one or more paralogues retain the broader ancestral expression²⁹. To distinguish between these alternatives, we analysed a subset of multi-gene families in which both the single amphioxus orthologue and the union of the vertebrate ohnologues—and thus probably the ancestral gene—were expressed across all nine samples that we compared. We then identified (i) gene families in which all vertebrate paralogues were expressed in all domains (termed 'redundancy'), (ii) gene families in which none of the vertebrate members had expression across all domains (termed 'subfunctionalization')²⁷ and (iii) gene families in which one or more vertebrate ohnologues were expressed in all domains, but at least one ohnologue was not (termed 'specialization') (Fig. 6c). We obtained very similar results for the three vertebrate species we studied (Fig. 6d): between 80 and 88% of gene families were subfunctionalized or specialized, which implies that ancestral expression domains have been lost in at least one member. Moreover, specialization was consistently more frequent than subfunctionalization as a fate for ohnologues with broad ancestral expression.

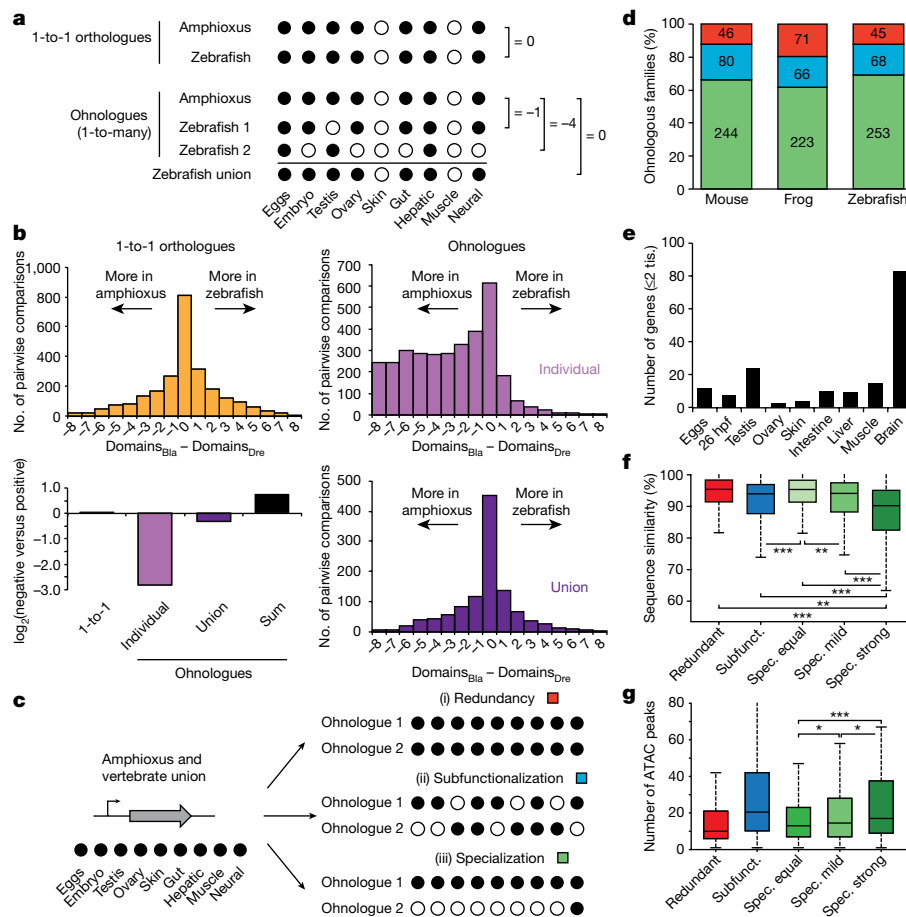


Fig. 6 | Expression specialization is the main fate after WGD.

a, Schematic of the analysis shown in **b**. Expression is binarized for each gene across the nine homologous samples ('on', black dots; normalized cRPKM > 5). **b**, Distribution of the difference in positive domains between zebrafish (domains_{Dre}) and amphioxus (domains_{Bla}) for 1-to-1 orthologues (2,478 gene pairs; yellow), individual ohnologues (3,427 gene pairs in 1,135 families; lilac) and the union of all vertebrate ohnologues in a family (purple). Bottom left, log₂ of the ratio between zebrafish genes with negative and positive score for each category. 'Sum' (black), binarization of family expression after summing the raw expression values for all ohnologues. **c**, Schematic of the analyses shown in **d**, representing the three possible fates after WGD. **d**, Distribution of fates after WGD for families of ohnologues. **e**, Number of ohnologues with strong

specialization in zebrafish expressed in each domain. Tis., tissue. **f**, Distribution of the percentage of nucleotide sequence similarity between human and mouse by family type. Ohnologues from specialized families are divided into 'spec. equal' (which maintain all expression domains), 'spec. mild' (which have lost some but maintained more than two expression domains) and 'spec. strong' (≤2 expression domains remain). Subfunct., subfunctionalized. **g**, Distribution of the number of APREs within GREAT regions for zebrafish ohnologues for each category. Only statistical comparisons within specialized families are shown. *P* values in **f** and **g** correspond to two- and one-sided Wilcoxon sum-rank tests between the indicated groups, respectively. *0.05 > *P* value ≥ 0.01, **0.01 > *P* value ≥ 0.001, ****P* value < 0.001. Exact *P* values and sample sizes are provided in Supplementary Data 2, dataset 8.

Ohnologues that have experienced strong specialization (≤2 remaining expression domains) retained expression more often in neural tissues (Fig. 6e and Extended Data Fig. 10f–i) and were generally not expressed in additional vertebrate-specific tissues (Supplementary Information). Furthermore, they showed the fastest rates of sequence evolution (Fig. 6f and Extended Data Fig. 10j–l), consistent with an optimization of their coding sequence to perform their function in a specific tissue and/or with the evolution of novel functions (neofunctionalization). Ohnologues from specialized families that have lost expression domains showed significantly more associated APREs than ohnologues with the full ancestral expression (Fig. 6g). We observed a strong positive relationship between the number of ancestral expression domains lost and the number of APREs associated with specialized ohnologues (Extended Data Fig. 10m). This implies that the specialization of gene expression after WGD does not occur primarily through loss of ancestral tissue-specific enhancers, but rather by a complex remodelling of regulatory landscapes that involves recruitment of novel, tissue-specific regulatory elements.

Discussion

By applying functional genomics approaches to the cephalochordate amphioxus, we have deepened our understanding of the origin and

evolution of chordate genomes. We identified APREs in amphioxus, the activation of which is tightly associated with differential DNA demethylation in adult tissues—a mechanism previously thought to be specific to vertebrates. Additional cases may be subsequently found in other non-vertebrate species when similar multi-omics datasets are analysed. In amphioxus, APREs of this type usually fall within gene bodies of widely expressed genes, which suggests that gene regulation by demethylation could have originated as a mechanism to allow better definition of enhancers in a hyper-methylated intra-genic context. If so, this mechanism could have been co-opted into new genomic contexts—that is, distal intergenic enhancers—later in the evolution of vertebrate genomes, which are characterized by their pervasive, genome-wide hypermethylation.

We also found a consistently higher number of open chromatin regions per gene in vertebrates than in amphioxus. This pattern is observed at a genome-wide level, but is particularly evident for distal APREs and in gene families that retain multiple ohnologues after WGD; these families are enriched for regulatory genes with large regulatory landscapes. Finally, we detected a large degree of specialization in expression for retained ohnologues, with the vast majority of multi-gene families with broad ancestral expression having at least one member

that restricted its expression breadth. Through this mechanism, vertebrates have increased their repertoire of tightly regulated genes, which has potentially contributed to tissue-specific evolution. Gene-expression specialization was accompanied by faster evolution of protein-coding sequences, and by an increase—rather than a decrease—in the number of regulatory elements. Taken together, these observations indicate that the two rounds of WGD not only caused an expansion and diversification of gene repertoires in vertebrates, but also allowed functional and expression specialization of the extra copies by increasing the complexity of their gene regulatory landscapes. We suggest that these changes to the gene regulatory landscapes underpinned the evolution of morphological specializations in vertebrates.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0734-6>.

Received: 22 November 2017; Accepted: 18 October 2018;

Published online 21 November 2018.

- Bertrand, S. & Escriva, H. Evolutionary crossroads in developmental biology: amphioxus. *Development* **138**, 4819–4830 (2011).
- Dehal, P. & Boore, J. L. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* **3**, e314 (2005).
- Putnam, N. H. et al. The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064–1071 (2008).
- Holland, L. Z. et al. The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Res.* **18**, 1100–1111 (2008).
- International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- Nelson, C. E., Hersh, B. M. & Carroll, S. B. The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol.* **5**, R25 (2004).
- Bogdanović, O. et al. Active DNA demethylation at enhancers during the vertebrate phylotypic period. *Nat. Genet.* **48**, 417–426 (2016).
- Berthelot, C., Villar, D., Horvath, J. E., Odom, D. T. & Flicek, P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nat. Ecol. Evol.* **2**, 152–163 (2018).
- Reilly, S. K. et al. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347**, 1155–1159 (2015).
- Villar, D. et al. Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).
- Vierstra, J. et al. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* **346**, 1007–1012 (2014).
- Boyle, A. P. et al. Comparative analysis of regulatory information and circuits across distant species. *Nature* **512**, 453–456 (2014).
- Gerstein, M. B. et al. Comparative analysis of the transcriptome across distant species. *Nature* **512**, 445–448 (2014).
- Hendrich, B. & Tweedie, S. The methyl-CpG binding domain and the evolving role of DNA methylation in animals. *Trends Genet.* **19**, 269–277 (2003).
- Irimia, M. et al. Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. *Genome Res.* **22**, 2356–2367 (2012).
- Simakov, O. et al. Insights into bilaterian evolution from three spiral genomes. *Nature* **493**, 526–531 (2013).
- Wang, X. et al. Genome-wide and single-base resolution DNA methylomes of the Pacific oyster *Crassostrea gigas* provide insight into the evolution of invertebrate CpG methylation. *BMC Genomics* **15**, 1119 (2014).
- Albalat, R., Martí-Solans, J. & Cañestro, C. DNA methylation in amphioxus: from ancestral functions to new roles in vertebrates. *Brief. Funct. Genomics* **11**, 142–155 (2012).
- Huang, S. et al. Decelerated genome evolution in modern vertebrates revealed by analysis of multiple lancelet genomes. *Nat. Commun.* **5**, 5896 (2014).
- Zhang, Y. et al. Nucleation of DNA repair factors by FOXA1 links DNA demethylation to transcriptional pioneering. *Nat. Genet.* **48**, 1003–1013 (2016).
- Irie, N. & Kuratani, S. Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nat. Commun.* **2**, 248 (2011).
- Hu, H. et al. Constrained vertebrate evolution by pleiotropic genes. *Nat. Ecol. Evol.* **1**, 1722–1730 (2017).
- Duboule, D. Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Development* **1994 Suppl.**, 135–142 (1994).
- Bogdanović, O. et al. Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. *Genome Res.* **22**, 2043–2053 (2012).
- Yue, F. et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).
- McLean, C. Y. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
- Force, A. et al. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).
- Yanai, I. et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
- Sandve, S. R., Rohlfs, R. V. & Hvidsten, T. R. Subfunctionalization versus neofunctionalization after whole-genome duplication. *Nat. Genet.* **50**, 908–909 (2018).

Acknowledgements This research was funded primarily by the European Research Council (ERC) under the European Union's Horizon 2020 and Seventh Framework Program FP7 research and innovation programs (ERC-AdG-LS8-740041 to J.L.G.-S., ERC-StG-LS2-637591 to M.I., a Marie Skłodowska-Curie Grant (658521) to I.M. and a FP7/2007-2013-ERC-268513 to P.W.H.H.), the Spanish Ministerio de Economía y Competitividad (BFU2016-74961-P to J.L.G.-S., RYC-2016-20089 to I.M., BFU2014-55076-P and BFU2017-89201-P to M.I. and BFU2014-55738-REDT to J.L.G.-S., M.I. and J.R.M.-M.), the 'Centro de Excelencia Severo Ochoa 2013-2017' (SEV-2012-0208), the 'Unidad de Excelencia María de Maetzu 2017-2021' (MDM-2016-0687), the People Program (Marie Curie Actions) of the European Union's Seventh Framework Program FP7 under REA grant agreement number 607142 (DevCom) to J.L.G.-S., and the CNRS and the ANR (ANR16-CE12-0008-01) to H.E. O.B. was supported by an Australian Research Council Discovery Early Career Researcher Award (DECRA; DE140101962). We acknowledge the support of the CERCA Programme/Generalitat de Catalunya and of the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) to the EMBL partnership. Additional sources of funding for all authors are listed in Supplementary Information.

Reviewer information *Nature* thanks D. Duboule and P. Flicek for their contribution to the peer review of this work.

Author contributions F.M., P.N.F., I.M., J.J.T., O.B., M.P., B.L., P.W.H.H., H.E., J.L.G.-S. and M.I. contributed to concept and study design. F.M., P.N.F., I.M., J.J.T., O.B., M.P., C.D.R.W., R.D.A., S.J.v.H., C.H.-U., K.S., Y.M., A. Louis, P.J.B., P.E.D., M.T.W., J.G.-F., R.L., B.L., P.W.H.H., J.L.G.-S. and M.I. performed computational analyses and data interpretation. O.B., E.d.I.C.-M., S.B., D.B., R.D.A., S.N., S.J.-G., D.A., L.B., J.P., B.A.-C., Y.L.P., A. Leon, L.S., E.F., P.C., J.R.M.-M., R.L., B.L., H.E., J.L.G.-S. and M.I. obtained biological material and generated next-generation sequencing data. I.M., J.J.T., E.d.I.C.-M., I.K., R.D.A., Z.K. and J.L.G.-S. performed transgenic assays. J.-M.A., S.M. and P.W. sequenced the genome. R.A., E.B.-G., C.C., F.C., S.D., D.E.K.F., S.H., V.L., G.A.B.M., P.P., M.S., H.S., I.S., T.T., O.M., A.X. and J.-K.Y. contributed to genome sequencing and gene family curation. I.M., H.E., J.L.G.-S. and M.I. coordinated the project. F.M., I.M., P.W.H.H. and M.I. wrote the main text, with input from all authors. Detailed contributions are listed in Supplementary Information. Animal illustrations by J.J.T., released under a Creative Commons Attribution (CC-BY) Licence.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0734-6>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0734-6>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to I.M., H.E., J.L.G. or M.I.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Animal husbandry and embryo staging. Amphioxus gametes were obtained by heat stimulation as previously described^{30,31}. Embryos were obtained by in vitro fertilization in filtered seawater and cultured at 19 °C. Staging was done based on previous publications^{32,33}; correspondence between developmental stages and hpf are provided in Supplementary Table 1. All protocols used for vertebrate species (zebrafish and medaka) have been approved by the Institutional Animal Care and Use Ethic Committee (PRBB–IACUEC, for CRG) or the Ethics Committee of the Andalusian Government (license numbers 450–1839 and 182–41106, for CABD–CSIC), and implemented according to national and European regulations. All experiments were carried out in accordance with the principles of the 3Rs (replacement, reduction and refinement).

Genome sequencing and assembly. Genomic DNA was extracted from a single *B. lanceolatum* adult male collected in Argeles-sur-Mer, France. The genome was sequenced using a combination of Illumina libraries from a range of inserts at Genoscope (897 million reads in total, with a paired-end coverage of 150×; Supplementary Table 2). A diploid assembly was generated using SOAPdenovo assembler³⁴ using a *k*-mer of 71. After gap closing, haplotypes were reconciled with Haplomerger³⁵.

Genome annotation. We generated deep coverage RNA-seq for 16 developmental stages and 9 adult tissues (4.16 billion reads in total). The bulk of strand-specific transcriptomic data was assembled de novo with Trinity³⁶, aligned and assembled into loci with the PASA pipeline³⁷. De novo gene models were built using Augustus³⁸ and subsequently refined with EVM³⁹ using PASA assemblies and aligned proteins from other species. In parallel, all strand-specific RNA-seq reads were mapped to the genome using Tophat⁴⁰, assembled using Cufflinks⁴¹ and open reading frames were predicted using Trans-decoder⁴². Models obtained using both these approaches were reconciled yielding a total 218,070 transcripts from 90,927 unified loci, of which 20,569 were protein-coding and had homologues in at least one of the other studied species (see ‘Comparative genomics’). Gene Ontology (GO) terms were assigned to amphioxus proteins based on their PFAM and Interpro domains, as well as blastp hits against human proteins (1×10^{-6}).

Repeats were annotated and filtered with RepeatMasker using a custom library generated with RepeatModeller. Long non-coding RNAs were identified by filtering all transcripts for protein-coding potential using CPAT⁴³ trained with zebrafish transcripts, and further discarding those that had a positive hit in a HMM search against the NR and PFAM databases (Extended Data Fig. 1g).

Comparative genomics. We used OMA⁴⁴ to reconstruct gene families and infer homology relationships based on well-established phylogenetic relationships between species⁴⁵, and further merged families sharing Ensembl paralogues with ‘Euteleostomi’ or ‘Vertebrata’ ancestry. To define the set of high-confidence ohnologue families (Supplementary Data 2, dataset 9), we retained families with two to four copies in three out of five vertebrates (excluding teleosts) and subjected them to phylogenetic reconciliation.

To assess genome sequence conservation, reciprocal whole-genome alignments of *Branchiostoma floridae*, *Branchiostoma belcheri* and *B. lanceolatum* were performed using LASTZ and processed with phastCons⁴⁶ to produce conservation scores. The distribution of phastCons scores in APREs was determined using ‘dynamic’ ATAC-seq peaks that showed no temporal discontinuity in activity.

Comparative transcriptomics. To investigate the evolutionary conservation of chordate development at the molecular level, newly generated data from zebrafish, medaka and amphioxus, as well as available data from the SRA (frog and chicken), were compared (Supplementary Data 2, dataset 3 and Supplementary Table 3). Gene expression was estimated with Kallisto⁴⁷ using Ensembl transcriptome annotations (Supplementary Table 4), and summing up transcripts per million (TPMs) from all transcript isoforms to obtain one individual gene-expression estimate per sample. We used single-copy orthologues to pair genes and used the Jensen–Shannon distance metrics after quantile normalization of TPMs to score distance between pairs of transcriptomes:

$$JSD_s = \sqrt{\frac{1}{2} \sum_{g=0}^{n_{og}} p_g \times \log \left(\frac{p_g}{\frac{1}{2}(p_g + q_g)} \right) + \frac{1}{2} \sum_{g=0}^{n_{og}} q_g \times \log \left(\frac{q_g}{\frac{1}{2}(p_g + q_g)} \right)}$$

Statistical robustness towards gene sampling was assessed by calculating transcriptomic distances based on 100 bootstrap replicates and estimating the standard deviation over these replicates.

To obtain groups of genes with similar dynamics of expression during development, genes were clustered based on their cRPKM⁴⁸ using the Mfuzz package⁴⁹. For this purpose, eight comparable stages were selected in amphioxus and zebrafish on the basis of conserved developmental landmarks such as fertilization,

gastrulation and organogenesis (Supplementary Table 5). The statistical significance of the orthologous gene overlap between pairs of clusters was assessed using upper-tail hypergeometric tests.

Modules of co-expressed genes across stages and adult tissues were inferred using WGCNA⁵⁰ with default parameters in amphioxus (17 samples) and zebrafish (27 samples) (Supplementary Table 6). The statistical significance of the orthologous gene overlap between pairs of clusters was assessed using upper-tail hypergeometric tests. The numbers of transcription-factor binding-site motifs detected in APREs in the basal regions of genes from any given cluster were standardized using *z*-scores.

To have a general assessment of the extent of conservation or divergence in gene expression among chordates at adult stages, we used neighbourhood analysis of conserved co-expression (NACC)²⁵, a method developed to compare heterogeneous, non-matched sample sets across species. NACC relies on comparisons of average distances between pairs of orthologous (genes A and B), the 20 genes with the closest transcriptomic distance (\bar{A} and \bar{B}) and their reciprocal orthologues in the other species (\bar{AB} and \bar{BA}), and is calculated as follows:

$$NACC = \frac{1}{2} [(\bar{AB} - \bar{A}) + (\bar{BA} - \bar{B})]$$

NACC calculations were performed for each family that contained a single amphioxus member and up to eight members in zebrafish and were also performed with randomized orthology relationships as a control.

Regulatory profiling. ATAC-seq. For amphioxus, medaka and zebrafish, ATAC-seq was performed in two biological replicates by directly transferring embryos in the lysis buffer, following the original protocol^{51,52}. ATAC-seq libraries were sequenced to produce an average of 66, 83 and 78 million reads for amphioxus, zebrafish and medaka, respectively. Reads were mapped with Bowtie2 and nucleosome-free pairs (insert < 120 bp) retained for peak-calling using MACS2⁵³, and the irreducible discovery rate was used to assess replicability. Nucleosome positioning was calculated from aligned ATAC-seq data using NucleoATAC⁵⁴. **Chromatin immunoprecipitation with sequencing (ChIP-seq).** Embryos of undetermined gender were fixed in 2% formaldehyde and ChIP was performed as previously described for other species⁵⁵. Chromatin was sonicated and incubated with the corresponding antibody (H3K4me3: ab8580, H3K27ac: ab4729 and H3K27me3: ab6002, from Abcam). An average of 30 million reads per library was generated. Reads were mapped with Bowtie2 and peaks called with MACS2⁵³, assuming default parameters.

4C-seq. Embryos of undetermined gender were fixed in 2% formaldehyde and chromatin was digested with DpnII and Csp6. Specific primers targeted the TSSs of the studied genes and included Illumina adapters. An average 5 million reads were generated for each of the two biological replicates. After mapping, reads were normalized per digestion fragment cut and interactions were identified using peakC⁵⁶ with low-coverage regions excluded.

MethylC-seq and RRBS. Genomic DNA was extracted as previously described⁵⁷, sonicated, purified and end-repaired. Bisulfite conversion was performed with the MethylCode Bisulfite Conversion Kit (Thermo Fisher Scientific). After Illumina library construction, an average of 73 million reads per sample were sequenced. RRBS libraries were prepared similarly to those for MethylC-seq, but with restriction digestion with MspI instead of sonication and PCR amplification. An average of 46 million reads per sample was generated. Reads were mapped to an in silico, bisulfite-converted *B. lanceolatum* reference genome^{7,58}. Differentially methylated regions in the CpG context were identified as previously described⁷. Differential transcription-factor motif enrichment was obtained with DiffBind from Bioconductor.

CAGE-seq. Libraries were constructed using the non-amplifying non-tagging Illumina CAGE protocol⁵⁹. Mouse CAGE-seq data were obtained from FANTOM5⁶⁰. Reads were aligned using Bowtie. Nearby individual CAGE TSSs were combined using the distance-based clustering method in CAGER⁶¹ to produce tag clusters, which summarize expression at individual promoters. Tag clusters were clustered across samples to produce comparable promoter regions, referred to as ‘consensus clusters’. The consensus clusters were then grouped by expression patterns using a self-organizing map⁶². We investigated the relative presence and enrichment of the following features: TATA box, YY1 motif, GC and AT content, SS and WW dinucleotides, first exons and nucleosome positioning signal. Heat maps were plotted for visualization by scanning either for exact dinucleotide matches or for position weight matrix matches at 80% of the maximum score. Position weight matrices for TATA and YY1 were taken from the JASPAR vertebrate collection.

Cis-regulatory comparisons. Depending on the analysis, an APRE was associated with a specific gene if it was located within: (i) the ‘basal’ region of the gene (–5 kb to +1 kb of the TSS; for comparisons of enriched motif composition) or (ii) the GREAT region of the gene (up to ±1 Mb of the TSS unless another basal region was found; for comparing the number of APREs per gene)²⁶. Stratification of gene

sets by GREAT or intergenic-region size between amphioxus and zebrafish was done using the function stratify from the matt suite⁶³, with a range of ± 500 bp.

The DNA-binding specificity of each transcription factor was predicted on the basis of the binding domain similarity to other transcription-factor family members, as previously performed⁶⁴. Transcription-factor motifs from CIS-BP version 1.02⁶⁴ were downloaded and clustered using GimmeMotifs⁶⁵ ($P \leq 0.0001$). Two hundred and forty-two clusters of motifs were assigned to one or more orthologous groups in both amphioxus and zebrafish and used for all analyses (Supplementary Data 2, dataset 10). These motifs were detected in APREs using the tools gimme threshold and gimme scan from GimmeMotifs⁶⁵.

Effect of WGDs on gene expression. Gene expression was binarized (1 if the normalized cRPKM > 5 , and 0 otherwise) across nine comparable samples in amphioxus and three vertebrate species (mouse, frog and zebrafish) (Supplementary Table 7). Then, for each amphioxus gene and vertebrate orthologue, the expression bias was measured by subtracting the number of positive-expression domains in amphioxus from that of vertebrates (Fig. 6a). The amphioxus gene-expression pattern was also compared to the union of the orthologues, as well as the pattern after binarizing the expression for the sum of cRPKM values of all family members. The analysis was restricted to families with a single member in amphioxus.

Next, we selected those ohnologue families for which the ancestral expression included the nine studied domains, as inferred from having expression in the single amphioxus orthologue and in the union of the family. For each gene family, we then defined (Fig. 6c): (i) redundancy (all vertebrate paralogues were expressed in all domains), (ii) subfunctionalization (none of the vertebrate members had expression across all domains²⁷), and (iii) specialization (one or more vertebrate ohnologues were expressed in all domains, but at least one ohnologue was not). Members of the later type were subdivided into 'strong' and 'mild' specialization if they retained ≤ 2 or more expression domains. We examined the transcript sequence similarity as well as the dN/dS between human and mouse (retrieved from Biomart), and the number of APREs associated with genes from different categories. Finally, we computed the τ tissue-specificity index as previously described²⁸, to assess more broadly the tissue specificity of ohnologues.

Transgenic assays in zebrafish and amphioxus. Enhancer reporter assays in zebrafish embryos were performed as previously described⁶⁶. Selected peaks were first amplified, cloned into a PCR8/GW/TOPO vector and transferred into a detection vector (including a *gata2* minimal promoter, a GFP reporter gene and a strong midbrain enhancer (z48) as an internal control)⁶⁷. Transgenic embryos were generated using the Tol2 transposon and transposase method⁶⁸. Three or more independent stable transgenic lines were generated for each construct as reported in Supplementary Table 8. For amphioxus reporter assays, selected peaks were amplified and transferred into a detection vector (including the *Branchiostoma* minimal actin promoter, a GFP reporter gene and piggyBac terminal repeats). Transgenic embryos were generated by the piggyBac transposase method.

In situ hybridization. Gene fragments that were synthetically designed or amplified by PCR from cDNA were sub-cloned into pBluescript II SK and used as templates for probe synthesis using the DIG labelling kit (Roche) and T3 RNA polymerase. Embryos at different developmental stages were fixed in PFA 4% dissolved in MOPS-EGTA buffer and in situ hybridization carried out as previously described⁶⁹, using BCIP/NBT as a chromogenic substrate.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. Custom code is available at <https://gitlab.com/groups/FunctionalAmphioxus>.

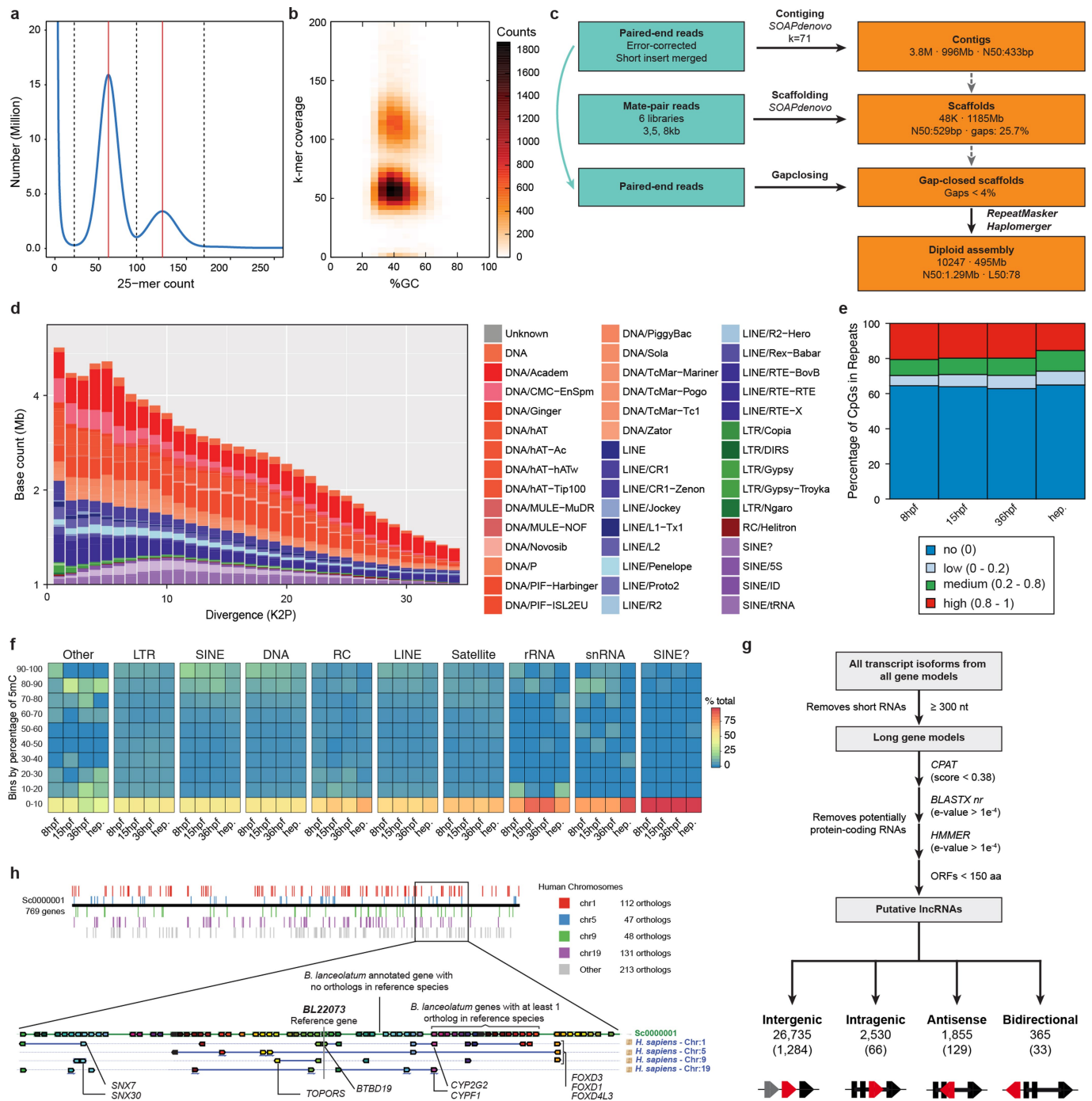
Data availability

Next-generation sequencing data have been deposited in Gene Expression Omnibus (GEO) under the following accession numbers: GSE106372 (ChIP-seq), GSE106428 (ATAC-seq), GSE106429 (CAGE-seq), GSE106430 (RNA-seq), GSE102144 (MethylC-seq and RRBS) and GSE115945 (4C-seq). Raw genome sequencing data and the genome assembly have been submitted to European Nucleotide Archive (ENA) under the accession number PRJEB13665. UCSC hub and annotation files are available at <http://amphiencode.github.io/>.

30. Fuentes, M. et al. Preliminary observations on the spawning conditions of the European amphioxus (*Branchiostoma lanceolatum*) in captivity. *J. Exp. Zool. B Mol. Dev. Evol.* **302B**, 384–391 (2004).
31. Fuentes, M. et al. Insights into spawning behavior and development of the European amphioxus (*Branchiostoma lanceolatum*). *J. Exp. Zool. B Mol. Dev. Evol.* **308B**, 484–493 (2007).
32. Hirakow, R. & Kajita, N. Electron microscopic study of the development of amphioxus, *Branchiostoma belcheri tsingtauense*: the gastrula. *J. Morphol.* **207**, 37–52 (1991).
33. Hirakow, R. & Kajita, N. Electron microscopic study of the development of amphioxus, *Branchiostoma belcheri tsingtauense*: the neurula and larva. *Kaibogaku Zasshi* **69**, 1–13 (1994).

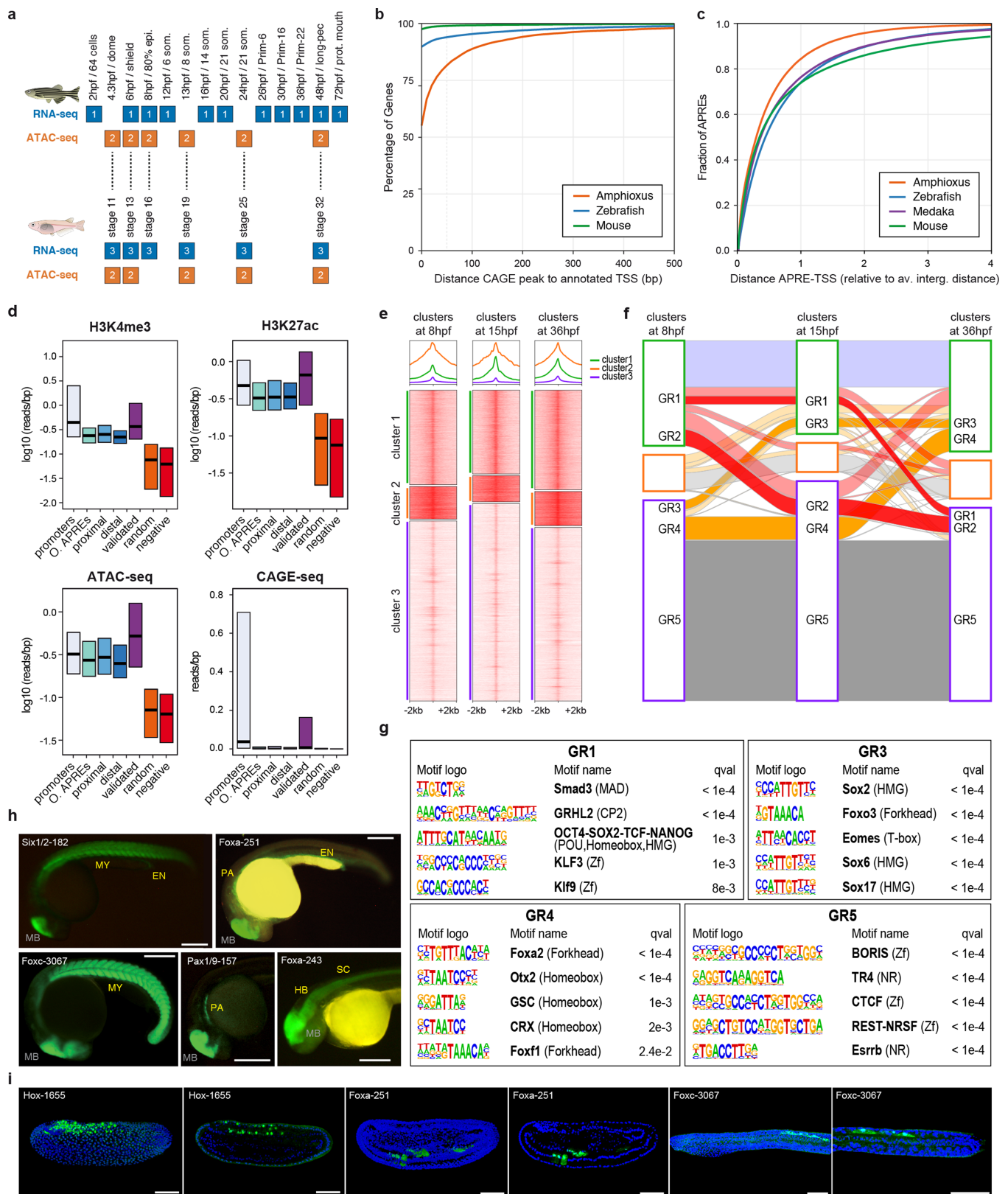
34. Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012).
35. Huang, S. et al. HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. *Genome Res.* **22**, 1581–1588 (2012).
36. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
37. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
38. Keller, O., Kollmar, M., Stanke, M. & Waack, S. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27**, 757–763 (2011).
39. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EvidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
40. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
41. Trapnell, C. et al. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
42. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protocols* **8**, 1494–1512 (2013).
43. Wang, L. et al. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41**, e74 (2013).
44. Roth, A. C., Gonnet, G. H. & Dessimoz, C. Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* **9**, 518 (2008).
45. Altenhoff, A. M., Gil, M., Gonnet, G. H. & Dessimoz, C. Inferring hierarchical orthologous groups from orthologous gene pairs. *PLoS ONE* **8**, e53786 (2013).
46. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
47. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
48. Labbé, R. M. et al. A comparative transcriptomic analysis reveals conserved features of stem cell pluripotency in planarians and mammals. *Stem Cells* **30**, 1734–1745 (2012).
49. Kumar, L. & Futschik, M. E. Mfuzz: a software package for soft clustering of microarray data. *Bioinformatics* **2**, 5–7 (2007).
50. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
51. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
52. Fernández-Miñán, A., Bessa, J., Tena, J. J. & Gómez-Skarmeta, J. L. Assay for transposase-accessible chromatin and circularized chromosome conformation capture, two methods to explore the regulatory landscapes of genes in zebrafish. *Methods Cell Biol.* **135**, 413–430 (2016).
53. Zhang, Y. et al. Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9**, R137 (2008).
54. Schep, A. N. et al. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res.* **25**, 1757–1770 (2015).
55. Bogdanović, O., Fernández-Miñán, A., Tena, J. J., de la Calle-Mustienes, E. & Gómez-Skarmeta, J. L. The developmental epigenomics toolbox: ChIP-seq and MethylCap-seq profiling of early zebrafish embryos. *Methods* **62**, 207–215 (2013).
56. Geeven, G., Teunissen, H., de Laat, W. & de Wit, E. peakC: a flexible, non-parametric peak calling package for 4C and Capture-C data. *Nucleic Acids Res.* **46**, e91 (2018).
57. Bogdanović, O. & Veenstra, G. J. Affinity-based enrichment strategies to assay methyl-CpG binding activity and DNA methylation in early *Xenopus* embryos. *BMC Res. Notes* **4**, 300 (2011).
58. Lister, R. et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471**, 68–73 (2011).
59. Murata, M. et al. Detecting expressed genes using CAGE. *Methods Mol. Biol.* **1164**, 67–85 (2014).
60. The FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
61. Haberle, V., Forrest, A. R., Hayashizaki, Y., Carninci, P. & Lenhard, B. CAGER: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res.* **43**, e51 (2015).
62. Wehrens, R. & Buydens, L. M. C. Self- and super-organising maps in R: the kohonen package. *J. Stat. Softw.* **21**, 1–19 (2007).
63. Gohr, A. & Irimia, M. Matt: Unix tools for alternative splicing analysis. *Bioinformatics* (2018).
64. Weirauch, M. T. et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
65. van Heeringen, S. J. & Veenstra, G. J. GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments. *Bioinformatics* **27**, 270–271 (2011).
66. Bessa, J. et al. Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. *Dev. Dyn.* **238**, 2409–2417 (2009).
67. Gehrke, A. R. et al. Deep conservation of wrist and digit enhancers in fish. *Proc. Natl Acad. Sci. USA* **112**, 803–808 (2015).

68. Kawakami, K. Transgenesis and gene trap methods in zebrafish by using the *Tol2* transposable element. *Methods Cell Biol.* **77**, 201–222 (2004).
69. Somorjai, I., Bertrand, S., Camasses, A., Haguénauer, A. & Escriva, H. Evidence for stasis and not genetic piracy in developmental expression patterns of *Branchiostoma lanceolatum* and *Branchiostoma floridae*, two amphioxus species that have evolved independently over the course of 200 Myr. *Dev. Genes Evol.* **218**, 703–713 (2008).
70. Tena, J. J. et al. Comparative epigenomics in distantly related teleost species identifies conserved *cis*-regulatory nodes active during the vertebrate phylotypic period. *Genome Res.* **24**, 1075–1085 (2014).
71. Acemel, R. D. et al. A single three-dimensional chromatin compartment in amphioxus indicates a stepwise evolution of vertebrate Hox bimodal regulation. *Nat. Genet.* **48**, 336–341 (2016).



Extended Data Fig. 1 | Summary of genomic assembly and repeat annotation. **a**, Spectrum of 25-mers in Illumina sequencing data that shows the bimodal distribution that is characteristic of highly polymorphic species. **b**, Heat map showing *k*-mer decomposition (y axis) across GC content (x axis). Both peaks show comparable GC content, which is consistent with them representing haploid versus diploid *k*-mers. **c**, Flow chart of the steps followed to obtain the *B. lanceolatum* assembly. **d**, Repeat landscape and its evolutionary history, shown by the proportion of repetitive elements with a given divergence (K2P) for their consensus in the repeat library (repeatScout). **e**, Percentage of methylated CpG dinucleotides within repetitive elements, at three developmental stages and in the adult hepatic diverticulum. **f**, Distribution of average levels of 5mC of different repeat families. Colour key indicates the percentage of repeats in each family with corresponding levels of average methylation. **g**, Computational pipeline to identify long non-coding RNAs (lncRNAs). Categories: antisense, lncRNA overlaps with a protein-coding gene in the reverse strand; intragenic, lncRNA overlaps with a protein-coding gene in the same strand; bidirectional, within 1 kbp of a TSS of a protein-coding gene in the antisense strand, probably a product of a

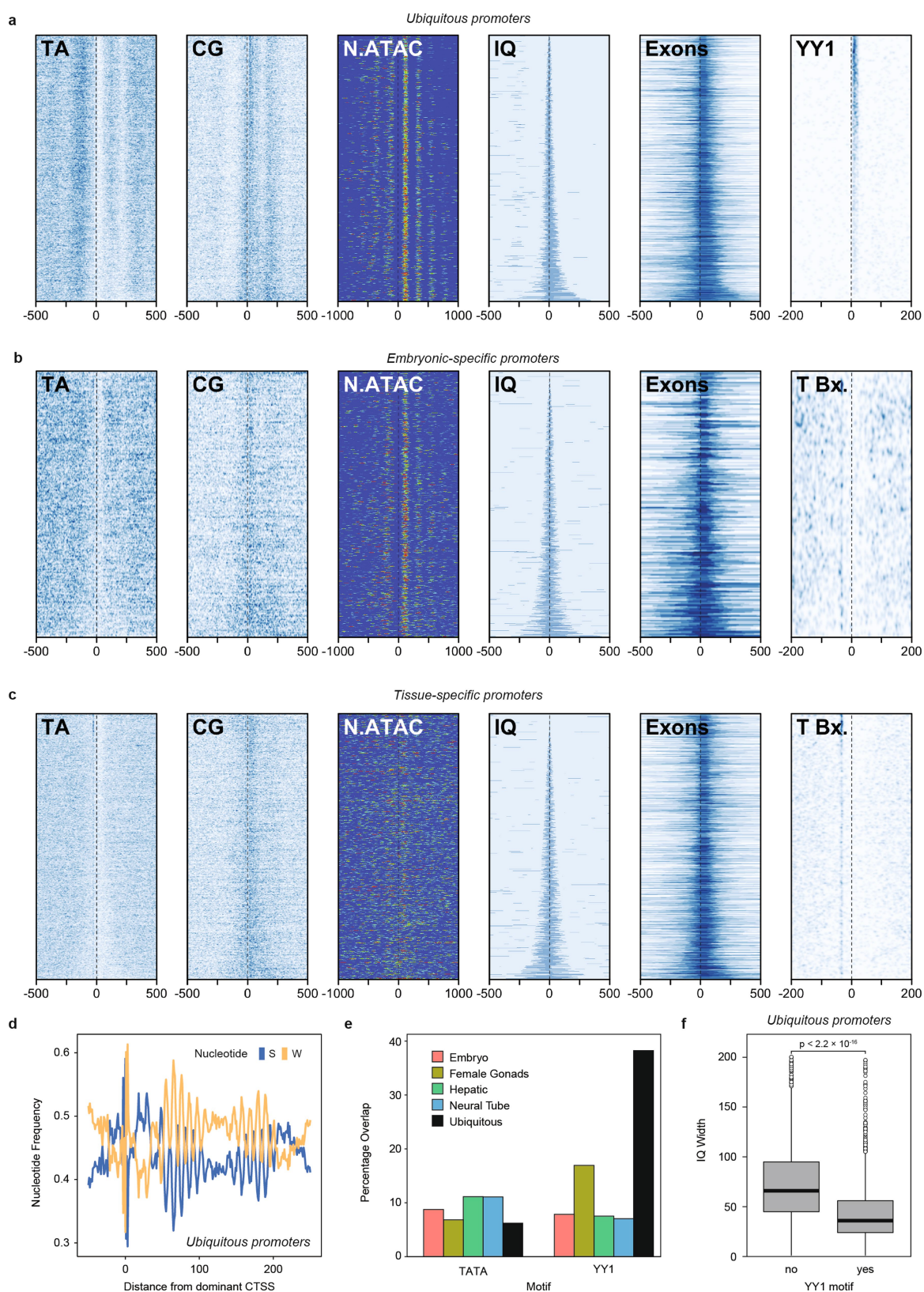
bidirectional promoter; intergenic, lncRNA does not overlap with any protein-coding gene. The total number in each category is indicated, with the number of those that are multi-exonic in parentheses. **h**, Quadruple conserved synteny between amphioxus and human. Top, amphioxus scaffold Sc00000001 aligned against the four human chromosomes with which it shares the highest number of orthologues (chr1, chr5, chr9 and chr19). In this scaffold, 277 out of 551 genes have clear orthologues in human, and 203 of these have orthologues on at least one of the four mentioned chromosomes. The black horizontal line represents the amphioxus scaffold, and each vertical coloured box an orthologous gene on the corresponding human chromosome. Bottom, modified view from Genomicus that is centred on the *BL22073* gene and spans Sc00000001: 7,736,434–8,850,041. On the top line, each amphioxus gene with at least one orthologue in the nine reference species is represented with an oriented coloured box. Human genes located in the four orthologous chromosomes are aligned underneath, in boxes of colours that correspond to those of their amphioxus pro-orthologues. The Genomicus server dedicated to amphioxus can be accessed at <http://genomicus.biologie.ens.fr/genomicus-amphioxus>.



Extended Data Fig. 2 | See next page for caption.

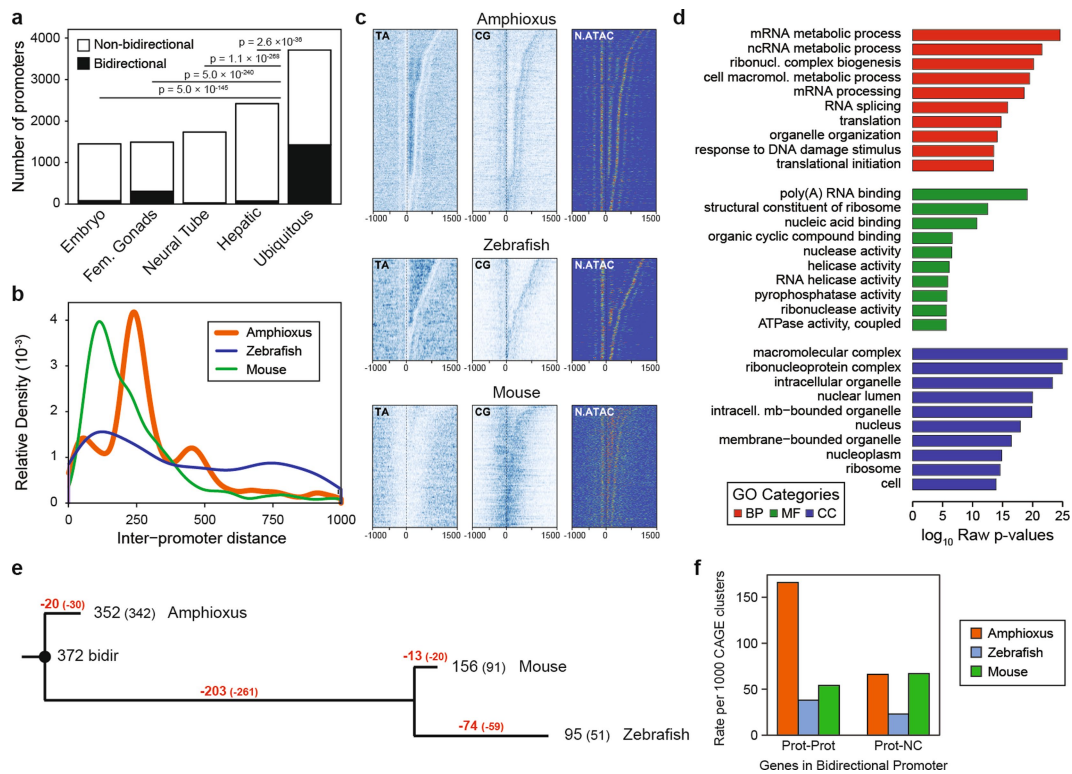
Extended Data Fig. 2 | Dynamics of chromatin marks on APREs and reporter assays. **a**, Summary of the zebrafish and medaka RNA-seq and ATAC-seq datasets generated for this study. Dashed lines indicate equivalent developmental stages in the two species, based on a previous study⁷⁰. The number of biological replicates is indicated for each experiment. Zebrafish 24-hpf ATAC-seq data are from a previous study⁶⁷. **b**, Cumulative distribution of the distance between CAGE-seq peaks and the closest annotated TSSs for genes with expression cRPKM > 5 in any of the samples covered by CAGE-seq (see Fig. 1a). Only CAGE-seq peaks within 1 kbp of an annotated TSS were tested (amphioxus: 10,435 peaks; zebrafish, 23,326 peaks; and mouse, 23,443 peaks). **c**, Cumulative distribution of distances between each APRE and the closest annotated TSS normalized by the average intergenic distance of the species (amphioxus, 83,471; zebrafish, 252,774; medaka, 174,139; and mouse, 216,857 APREs, as per Fig. 1c). **d**, Signal distribution of different marks within functional-genomic regions in amphioxus. \log_{10} of read counts of H3K4me3, H3K27ac and ATAC-seq, and raw read counts of CAGE-seq in promoters of homology-supported, protein-coding genes ($n = 26,501$), other APREs ('O. APREs', all APREs that do not overlap a TSS from any gene model; $n = 48,341$), proximal APREs ($n = 24,622$), distal APREs ($n = 11,881$), previously validated enhancers ($n = 43$; Supplementary Table 9), random regions ($n = 88,413$) and negative regions (excluding ATAC-seq peaks, $n = 88,413$). For region designation, see Fig. 1c. For clarity, whiskers and outliers are not displayed. **e**, *k*-means clustering of APREs based on H3K27ac signal in three developmental stages. Cluster 1 and 3 APREs were considered as active and inactive, respectively. Average H3K27ac profiles are represented in the top panels. The number of APREs per cluster and stage are provided in Supplementary Data 2, dataset 8. **f**, Alluvial plot that shows the dynamics of each APRE among

the clusters described in **e**. APREs that remained active (cluster 1 in all stages) along the three developmental stages are represented in blue, constitutively inactive APREs (cluster 3 in all stages) in dark grey and dynamic APREs in red or orange (if inactivated or activated, respectively, during development). Five groups of APREs of special interest are highlighted with stronger colours and named GR1–GR5. **g**, Representative enriched DNA motifs found in each of the groups described in **f**. GR1 APREs were enriched in early motifs (for example, Smad3 and Oct4, Sox2 and Nanog); GR3 APREs in motifs of transcription factors involved in the generation of the three germ layers (for example, Foxo3, Sox6 and Sox17); GR4 APREs in tissue-specific transcription factors (for example, Foxa2, Otx2 and Crx); and GR5 APREs in CTCF and CTCF-like (BORIS) motifs. *q* values as provided by Homer. **h**, Lateral views of embryos from stable transgenic zebrafish lines at 24 hpf (except for Foxa-243, at 48 hpf) showing GFP expression driven by the amphioxus APREs listed in Supplementary Table 8 and highlighted in Supplementary Fig. 1. The number of independent founders with the same expression were as follows: Six1/2-182 (5/5), Foxa-243 (3/3), Foxa-251 (4/4), FoxC-3067 (6/6) and Pax1/9-157 (3/3). Midbrain expression corresponds to the positive-control enhancer included in the reporter constructs. EN, endoderm; HB, hindbrain; MY, myotomes; PA, pharyngeal arch; SC, spinal cord. Scale bar, 250 μ m. **i**, Lateral views of transient transgenic amphioxus embryos, showing GFP expression driven by the APREs highlighted in Supplementary Fig. 1a, b (Foxa-251 ($n = 46$ out of 52) and Foxc-3067 ($n = 27$ out of 35), respectively) and in a previous study⁷¹ (Hox-1655, $n = 72$ out of 80). For each element, left panels correspond to 3D rendering from sub-stacks and right panels to *z*-stack sagittal sections. Scale bar, 50 μ m. Anterior is to the left and dorsal to the top.



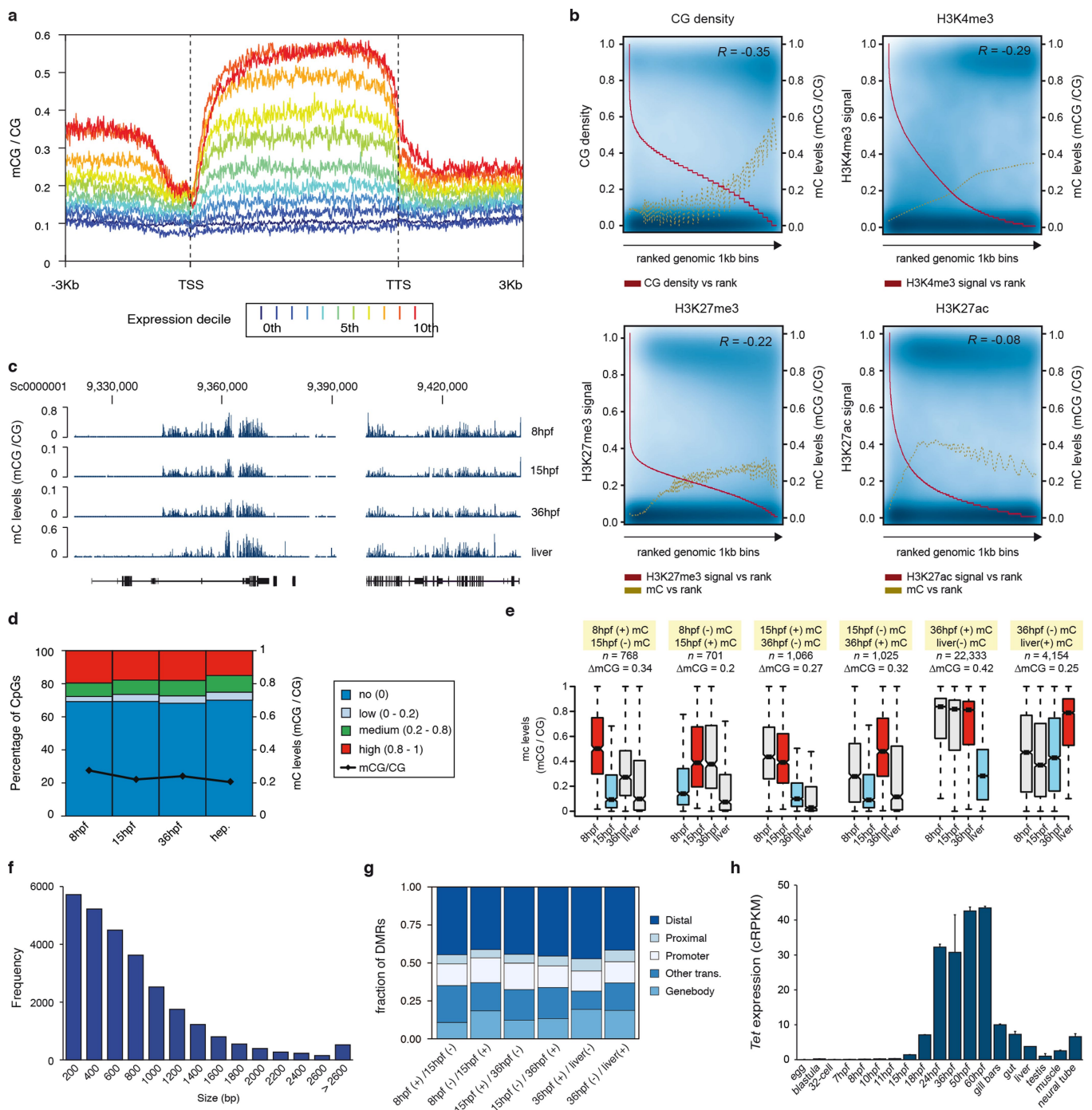
Extended Data Fig. 3 | Features of amphioxus promoters derived from CAGE-seq. **a–c**, Heat maps showing AT and CG signal, nucleosome positioning (derived from the NucleoATAC signal), promoter width (interquantile (IQ) range), first exon length and YY1 (**a**) or TATA box (**b**, **c**) motifs around ubiquitous (**a**, $n = 3,710$), embryonic-specific (**b**, $n = 1,451$) and tissue-specific (**c**, $n = 4,154$) promoters, sorted by promoter width. Position 0 corresponds to the main TSS. **d**, Ubiquitous promoters show strong evidence for a nucleosome positioned downstream

of the CAGE TSS, as judged from the 12-bp periodicity of W and S nucleotide density. **e**, Per cent of promoters of each category that have associated TATA box or YY1 motifs. Number of promoters: embryo, 1,451; female gonads, 1,494; hepatic, 2,420; neural tube, 1,734; and ubiquitous, 3,710. **f**, IQ width distribution of ubiquitous promoters ($n = 3,710$) with and without an associated YY1 motif. P value corresponds to two-sided Wilcoxon sum-rank tests.



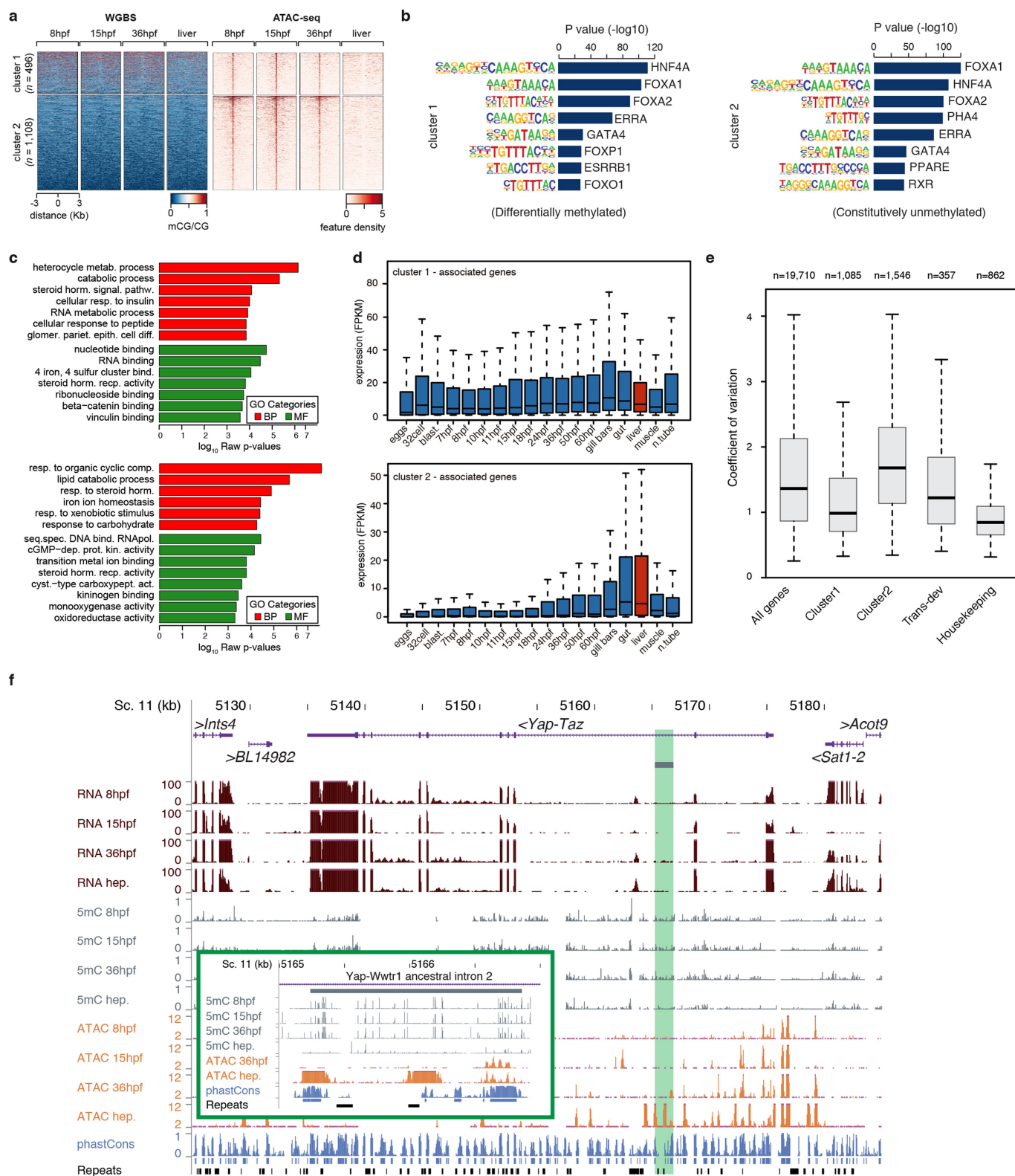
Extended Data Fig. 4 | Characteristics and evolution of bidirectional promoters. **a**, Number of bidirectional and non-bidirectional promoters identified for each regulatory category. P values correspond to two-sided Fisher's exact tests against ubiquitous promoters. **b**, Distribution of distance between bidirectional promoters in each species (amphioxus, 1,975; zebrafish, 549; and mouse, 876 pairs of promoters). The distance between amphioxus peaks closely corresponds to integral nucleosome spacing. **c**, Heat maps of TA, CG and nucleosome occupancy (derived from the NucleoATAC signal) around bidirectional promoter pairs in amphioxus ($n = 1,975$), mouse ($n = 876$) and zebrafish ($n = 549$), arranged by the distance between the two CAGE TSSs. In amphioxus, both TA and NucleoATAC signals indicate regions in which 0, 1 or 2 nucleosomes separate promoters. **d**, Enriched GO terms for genes associated with bidirectional promoters in amphioxus. Uncorrected P values correspond to two-sided Fisher's exact tests as provided by topGO.

e, Inferred evolutionary dynamics of 372 putatively ancestral bidirectional promoters among chordate groups. Red, number of inferred losses and disentanglements; black, number of detected bidirectional promoters by CAGE-seq (in brackets) or microsynteny (neighbouring genes in a 5' to 5' orientation) for each species. In parentheses, number of lost and disentangled (red) or retained (black) bidirectional promoters when considering only the cases supported by CAGE-seq. **f**, In vertebrates, disentanglement was not accompanied by a general increase in the fraction of bidirectional promoters with antisense non-coding transcription, as shown by the relative number of CAGE clusters identified as bidirectional promoters that are composed of two protein-coding genes ('Prot-Prot') or of one protein-coding and one non-coding or non-annotated locus ('Prot-NC'). The total number of uniquely annotated, protein-coding-associated CAGE promoters was amphioxus, 11,789; mouse, 13,654; and zebrafish, 14,014.



Extended Data Fig. 5 | 5mC dynamics in amphioxus. **a**, 5mC levels across gene bodies ($n = 20,569$) from different expression deciles (0th, not expressed; 10th, highest expression). TTS, transcription termination site. **b**, Scatter plots of levels of 5mC and CpG density, H3K4me3, H3K27me3 and H3K27ac in 1-kbp genomic bins sorted on the basis of feature rank. The red line tracks anti-correlation between feature density and rank number (a low rank number implies high feature density). The golden line represents a smoothing spline of 5mC signal versus feature rank number. Pearson correlation coefficients (R) are displayed in the top right corner of each panel. **c**, UCSC browser excerpt of 5mC patterns for selected regions. **d**, Percentage of methylated CpG dinucleotides in 8-hpf ($n = 19,657,388$), 15-hpf ($n = 21,247,615$), 36-hpf ($n = 21,702,000$) and hepatic (adult, $n = 19,240,245$) amphioxus samples. Black line indicates the fraction between methylated and non-methylated CpGs at each stage. **e**, Box

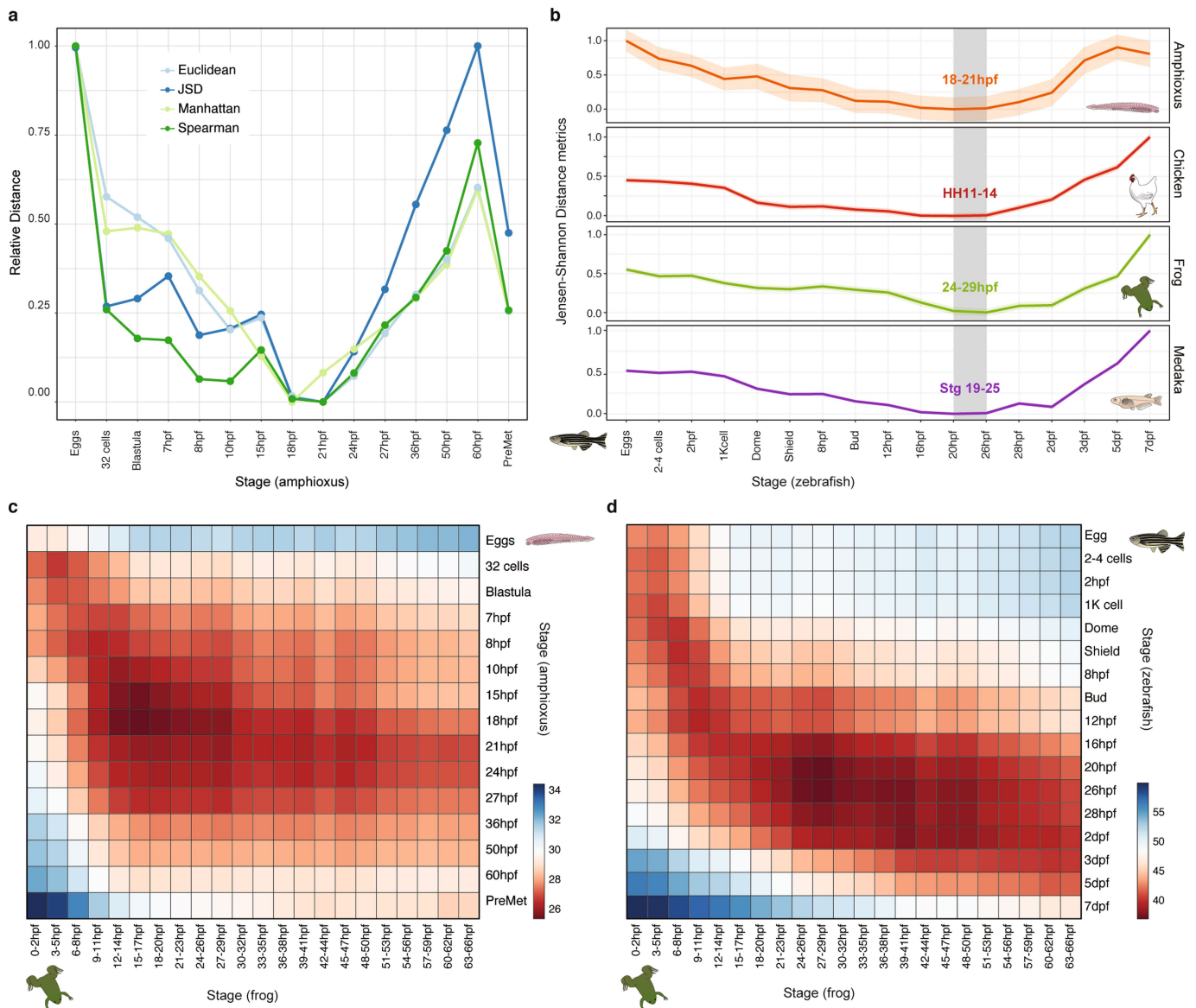
plots of average 5mC levels in different types of differentially methylated regions (DMRs) at each stage. Δ mCG denotes the change in the fraction of methylated CpGs between the two stages used for identification of DMRs (red (hyper) and blue (hypo) boxes). The number of DMRs were as follows: 8 hpf(+)–15 hpf(–), 768; 8 hpf(–)–15 hpf(+), 701; 15 hpf(+)-36 hpf(–), 1,066; 15 hpf(–)–36 hpf(+), 1,025; 36 hpf(+)-liver(–), 22,333; and 36 hpf(–)-liver(+), 4,154. The coordinates for all DMRs are provided in Supplementary Data 2, dataset 11. **f**, Distribution of DMR sizes (in bp). **g**, Genomic distribution of DMRs identified for each sample. ‘Other trans.’ DMRs that overlap with gene models that were not defined as being supported by orthology. **h**, Expression (cRPKM) of the amphioxus *Tet* orthologue in embryos and adult tissues. Error bars represent standard error of the mean (the number of replicates for each RNA-seq dataset is provided in Fig. 1a).



Extended Data Fig. 6 | See next page for caption.

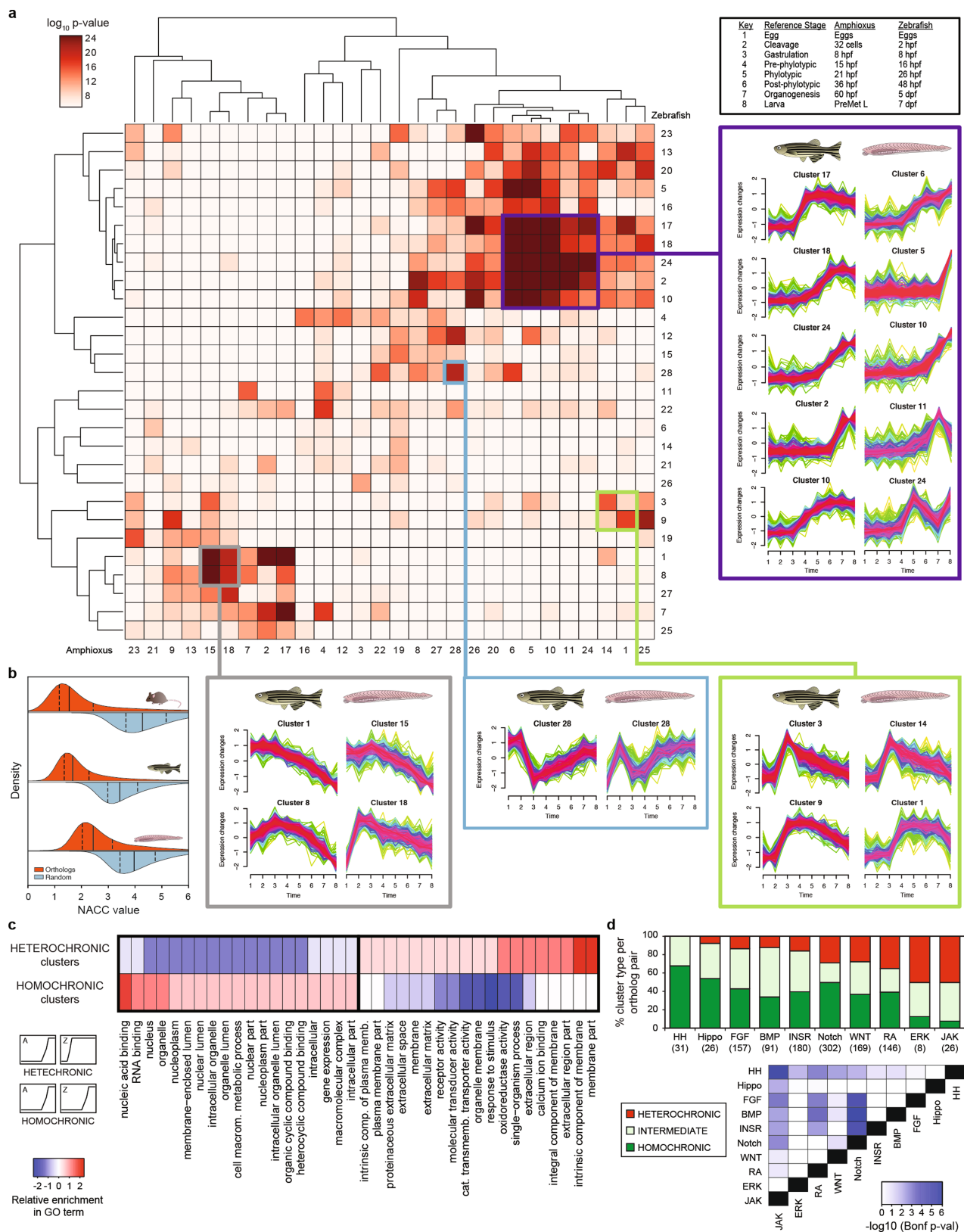
Extended Data Fig. 6 | Developmental 5mC dynamics at APREs in amphioxus. **a**, *k*-means clustering ($n = 2$) of 5mC signal over embryo-specific open-chromatin regions (that is, APREs), assessed by ATAC-seq (Supplementary Table 10). **b**, The most significantly enriched transcription-factor binding-site motifs in APREs that display different developmental 5mC patterns in Fig. 2b. Uncorrected *P* values as provided by MEME. All plotted motifs had Benjamini-corrected *q* values of 0. **c**, GO enrichment for genes associated with cluster 1 (top) or cluster 2 (bottom) APREs from Fig. 2b. Uncorrected *P* values correspond to two-sided Fisher's exact tests as calculated by topGO. **d**, Distribution of expression values (cRPKMs) across all samples for genes associated with cluster 1 (top, $n = 1,114$) or cluster 2 (bottom, $n = 1,594$) APREs from

Fig. 2b. **e**, Distribution of the coefficients of variation for genes associated with cluster 1 or cluster 2 APREs from Fig. 2b, as well as all ($n = 19,710$), trans-dev ($n = 357$) and house-keeping ($n = 862$) amphioxus genes. **f**, Example of a potentially conserved (zebrafish to amphioxus) DMR associated with *yap1*, a major transcription factor of the Hippo pathway. The inset corresponds to the region highlighted in green. The two orthologous genomic regions in zebrafish are shown in Supplementary Fig. 2. Additional cases included genes that contained APREs that are likely to regulate neighbouring liver-specific genes ('bystander' genes) (Supplementary Table 11). The number of replicates for each experiment displayed in each track is provided in Fig. 1a.



Extended Data Fig. 7 | Periods of maximal transcriptomic similarity across chordate development. **a**, Stages of minimal transcriptomic distance obtained in the comparison between amphioxus and zebrafish for four alternative distance methods (Euclidean, Manhattan and Jensen–Shannon distances, and Spearman correlation). Values are normalized to minimal (0) and maximal (1) for each metric. **b**, Stages of minimal transcriptomic divergence shown as the smallest Jensen–Shannon distance between zebrafish stages and four chordate species. The shaded area surrounding the line that connects the stages is the standard deviation, derived from 100 bootstrap replicates of the orthologous gene set.

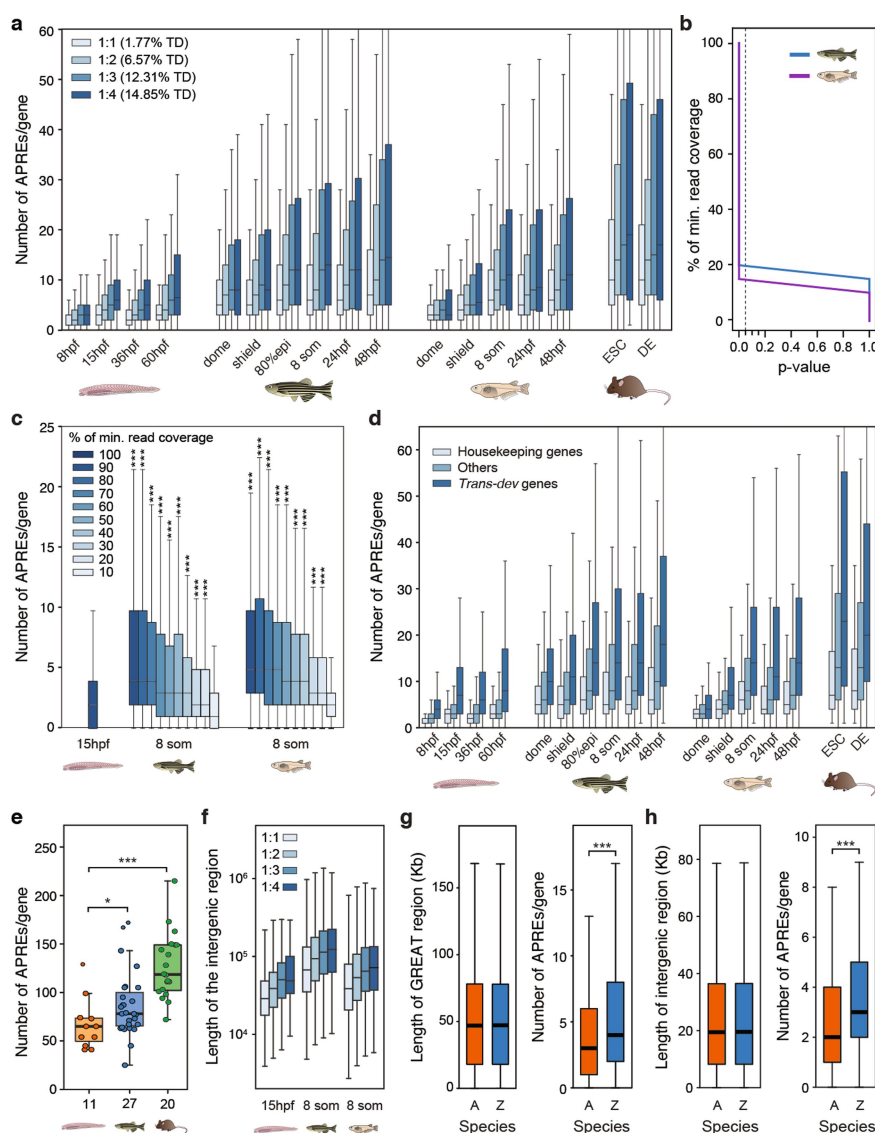
The grey box outlines the ‘phylotypic’ period of minimal divergence; the corresponding periods are indicated for each species as the range provided by the two closest stages. **c**, **d**, Heat maps of pairwise transcriptomic distances (Jensen–Shannon distance metric) between pairs of chordate species, amphioxus and frog (**c**), and zebrafish and frog (**d**). In both heat maps, the smallest distance (red) indicates maximal similarity of the transcriptome. The periods of minimal divergence of the transcriptome are earlier for the amphioxus–frog comparison than for the zebrafish–frog comparison.



Extended Data Fig. 8 | See next page for caption.

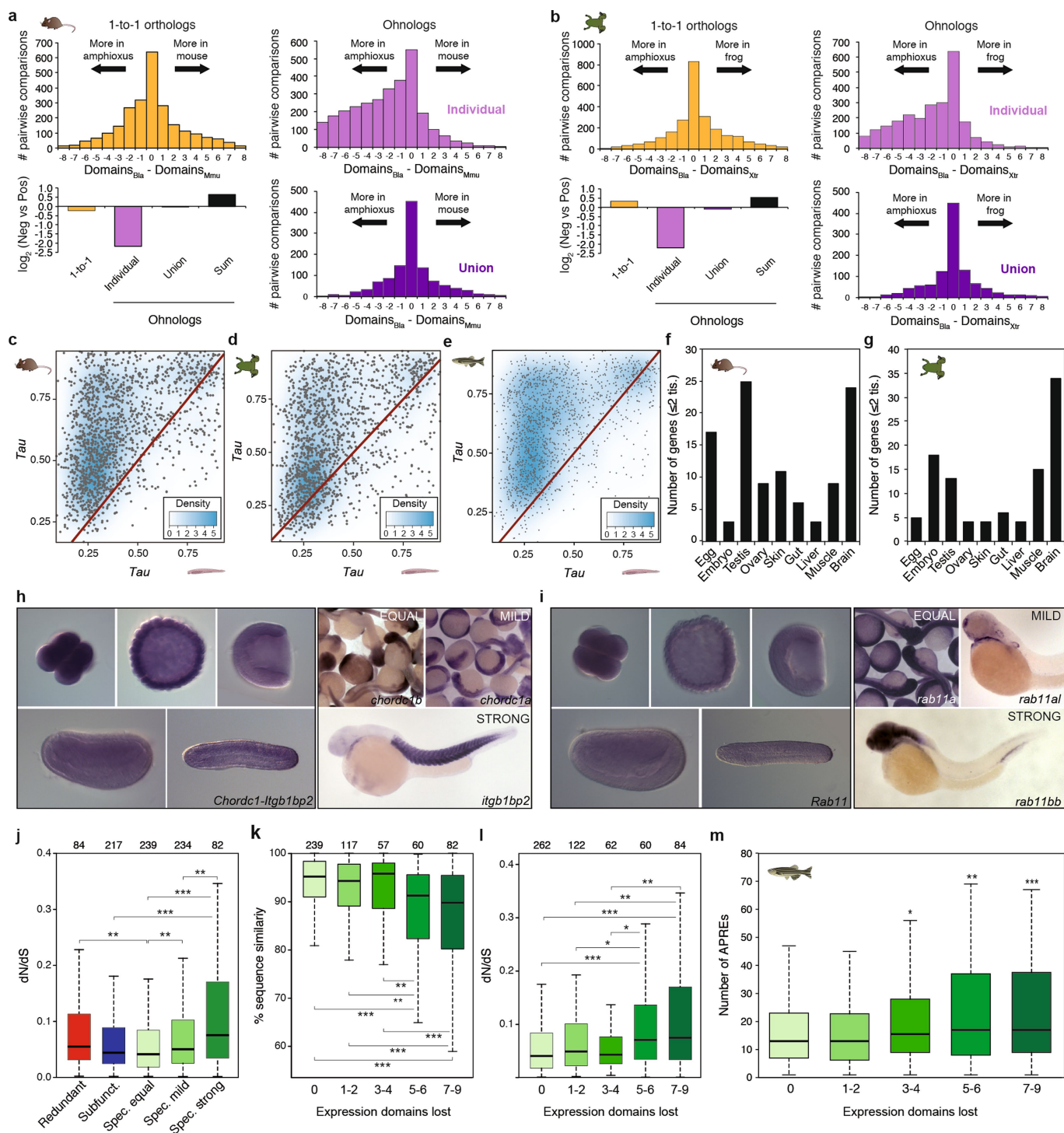
Extended Data Fig. 8 | Comparison of temporal gene expression profiles in amphioxus and zebrafish. **a**, Heat map showing the significance of orthologous gene overlap between Mfuzz clusters across eight matched developmental stages in amphioxus and zebrafish as derived from an upper-tail hypergeometric test. Some clusters with highly significant overlap are highlighted, and their corresponding temporal expression profiles are shown. The profiles of all clusters for the two species are included in Supplementary Figs. 3, 4. Exact P values and sample sizes are provided in Supplementary Data 2, dataset 8. **b**, Distributions of NACC values for orthologous genes (in red) or random orthology assignments (blue) for each species against human. Lower NACC values imply higher conservation of relative expression. Solid lines show the median, and the dashed lines mark the interquartile range. The number of orthologue pairs were as follows: mouse, 15,109; zebrafish, 16,480;

and amphioxus, 8,633. **c**, Differentially enriched GO terms among pairs of zebrafish and amphioxus Mfuzz clusters with significant orthologue overlap ($P < 10^{-10}$ upper-tail hypergeometric test) with homochronic (48 pairs) and heterochronic (35 pairs) patterns. The GO enrichment of a group was calculated as the number of cluster pairs with significant enrichment for that given term (Supplementary Data 2, dataset 12). **d**, Top, per cent of zebrafish genes from each developmental pathway we studied, based on the temporal similarity of their corresponding Mfuzz cluster (homochronic, heterochronic or intermediate). Only genes belonging to clusters with significant orthologue overlap were analysed; the number of genes is provided in parenthesis below the pathway name. Bottom, pairwise comparisons between developmental pathway distributions. P values correspond to Bonferroni-corrected, two-sided, three-way Fisher's exact tests.



Extended Data Fig. 9 | Higher regulatory content in vertebrate genomes. **a**, Distribution of the number of APREs per the regulatory landscape of a gene (as determined by GREAT²⁶), at different developmental stages or cell lines of four chordate species (amphioxus, zebrafish, medaka and mouse). Orthologous gene families are split according to the number of ohnologues that are retained per family (from 1 to 4, using mouse as a reference species for the ohnologue counts). The percentage of developmental regulatory genes (trans-dev, TD) in each category is indicated. **b**, *P* values of one-sided Mann–Whitney *U* tests against the amphioxus peak-number distribution using 100% of the minimum read coverage for different levels of down-sampling of the zebrafish and medaka samples. **c**, Distribution of the number of APREs in the GREAT region of the gene, called after down-sampling the reads of the two vertebrate samples to different fractions of the sample with the minimum effective coverage in our study (~21 reads per kbp for the 36-hpf sample in amphioxus). Asterisks correspond to the significance of the *P* values of Mann–Whitney *U* tests against the amphioxus peak-number distribution using 100% of the minimum-read coverage. The number of genes per box was as follows: amphioxus, 20,569; zebrafish,

20,053; and medaka, 15,978. **d**, As in **a**, but with gene families separated according to functional categories (housekeeping, trans-dev and others). **e**, Number of APREs per regulatory landscape determined using 4C-seq, for 58 members of 11 trans-dev families. The number of genes probed in each species is indicated on the x axis. **f**, Distribution of the length of the intergenic regions from the genes plotted in **a** for the indicated stages. **g**, Distributions of GREAT-region sizes (left) and number of APREs per gene (right) for a subset of 10,186 pairs of genes with matched GREAT-region size distributions (± 500 bp) in amphioxus and zebrafish. **h**, Distributions of intergenic-region sizes (left) and number of APREs per gene (right) for a subset of 13,941 pairs of genes with matched intergenic-region size distributions (± 500 bp) in amphioxus and zebrafish. *P* values correspond to Mann–Whitney *U* tests: *0.05 > *P* value \geq 0.01, **0.01 > *P* value \geq 0.001, ****P* value < 0.001. In **a** and **d**, all comparisons between each distribution of a vertebrate species and the equivalent distribution in amphioxus produced significant *P* values (*P* value < 0.001); for simplicity, in these panels asterisks are not shown. Exact *P* values and sample sizes are provided in Supplementary Data 2, dataset 8.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Regulatory evolution after vertebrate WGD.

a, b, For each mouse (**a**) or frog (**b**) gene, the number of positive-expression domains across nine equivalent samples is subtracted from the number of domains in which the single amphioxus orthologue is expressed. The distribution of the difference in domains between the amphioxus and the vertebrate species is plotted for 1-to-1 orthologues (2,450 and 2,484 gene pairs for mouse and frog, respectively; yellow), individual ohnologues (3,011 and 2,637 gene pairs in 1,212 and 1,094 families for mouse and frog, respectively; lilac) and the union of all vertebrate ohnologues in a family (purple). Bottom left, \log_2 of the ratio between the sum of all mouse (**a**) or frog (**b**) genes with negative versus positive score for each orthology group. 'Sum' (black), binarization of family expression is performed after summing the raw expression values for all ohnologues. **c–e**, Density scattered plot of the τ values for pairs of mouse (**c**, $n = 1,502$), frog (**d**, $n = 1,495$) and zebrafish (**e**, $n = 1,498$) and amphioxus orthologues from multi-gene families in vertebrates. **f, g**, Number of ohnologues with strong specialization (≤ 2 remaining expression domains) in mouse (**f**) or frog (**g**) expressed in each tissue or

developmental stage. **h, i**, Representative in situ hybridization assays in zebrafish embryos for different members of specialized families (right) and for the single amphioxus orthologue (left) (Chordc1 and Itgb1bp2 (**h**) and Rab11 (**i**)). Zebrafish image data for this paper were retrieved from the Zebrafish Information Network (ZFIN), University of Oregon, Eugene, OR 97403-5274; (<http://zfin.org/>, accessed May 2018) and are used with the permission of B. Thisse. Amphioxus in situ hybridization was performed once using 10 embryos per probe, all of which showed the same expression pattern. **j**, Distribution of the dN/dS ratio between human and mouse for different classes of ohnologues based on their fate after WGD. **k, l**, Distribution of the percentage of nucleotide sequence similarity (**k**) or dN/dS ratio (**l**) between human and mouse for ohnologues grouped by the number of expression domains lost. **m**, Distribution of the number of APREs within GREAT regions for zebrafish ohnologues grouped by the number of expression domains lost. *P* values in **j–m** correspond to Wilcoxon sum-rank tests. * $0.5 > P \text{ value} \geq 0.01$; ** $0.01 > P \text{ value} \geq 0.001$; *** $P \text{ value} < 0.001$.

PtdIns4P on dispersed *trans*-Golgi network mediates NLRP3 inflammasome activation

Jueqi Chen¹ & Zhijian J. Chen^{1,2*}

The NLRP3 inflammasome, which has been linked to human inflammatory diseases, is activated by diverse stimuli. How these stimuli activate NLRP3 is unknown. Here we show that different NLRP3 stimuli lead to disassembly of the *trans*-Golgi network (TGN). NLRP3 is recruited to the dispersed TGN (dTGN) through ionic bonding between its conserved polybasic region and negatively charged phosphatidylinositol-4-phosphate (PtdIns4P) on the dTGN. The dTGN then serves as a scaffold for NLRP3 aggregation into multiple puncta, leading to polymerization of the adaptor protein ASC, thereby activating the downstream signalling cascade. Disruption of the interaction between NLRP3 and PtdIns4P on the dTGN blocked NLRP3 aggregation and downstream signalling. These results indicate that recruitment of NLRP3 to dTGN is an early and common cellular event that leads to NLRP3 aggregation and activation in response to diverse stimuli.

Inflammasomes are multiprotein complexes that serve as a platform for caspase-1-dependent activation of pro-inflammatory cytokines, including interleukin-1 β (IL-1 β), as well as induction of a specific form of inflammatory cell death termed pyroptosis. The NLRP3 inflammasome is unusual in that it can be triggered by many different stimuli, such as nigericin (an antibiotic from *Streptomyces hygroscopicus*) or ATP released from damaged cells¹. Unregulated NLRP3 stimulation can lead to uncontrolled infection, autoimmune diseases, neurodegenerative diseases, metabolic disorders, and many other human diseases². Given the chemical and structural diversity of these stimuli, and the lack of evidence that NLRP3 interacts directly with any of these molecules, the mechanism by which NLRP3 is activated remains unknown. After stimulation, NLRP3 recruits the adaptor protein ASC (also known as PYCARD)³. ASC then undergoes prion-like polymerization, forming a single large spherical structure (called a ‘speck’) in the perinuclear region^{4,5}, before recruiting caspase-1 to activate the downstream signalling cascade¹.

Using biochemical, imaging and genetic approaches in both reconstituted systems and primary macrophages, we have identified a new common cellular signal downstream of diverse NLRP3 stimuli: the disassembly of the TGN into various dispersed structures, forming the dTGN. NLRP3 is then recruited to the dTGN through an interaction between a polybasic region on NLRP3 and the negatively charged PtdIns4P on the dTGN. NLRP3 on the dTGN forms multiple puncta that induce ASC polymerization, thereby activating the downstream signalling cascade.

An in vitro assay for detection of NLRP3 activity

We reconstituted the NLRP3 pathway in the HEK-293T cell line, which does not express endogenous NLRP3, ASC or caspase-1 (Extended Data Fig. 1a, b). We established a biochemical assay in this system, aiming to examine specifically the NLRP3 activation step (Fig. 1a). In brief, we set up HEK-293T cell lines stably expressing either NLRP3 (293 NLRP3) or ASC and caspase-1 (293 ASC-casp1). Extracts from stimulated 293 NLRP3 (‘activator’) cells were mixed with 293 ASC-casp1 (‘recipient’) cells that had been permeabilized with perfringolysin O (PFO),

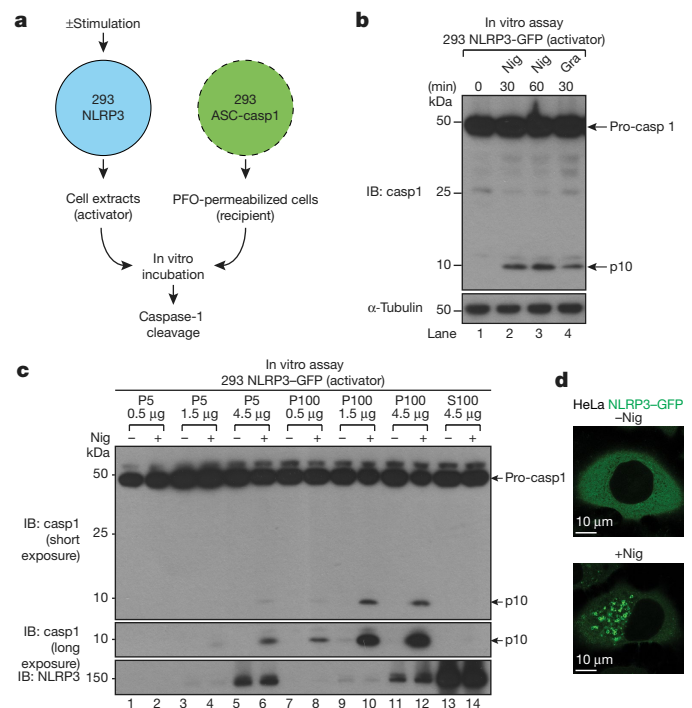


Fig. 1 | NLRP3 forms multiple puncta after stimulation. **a**, **b**, In vitro assay for examining NLRP3 activity. HEK-293T cells expressing NLRP3 were stimulated with nigericin (Nig) (10 μ M) or gramicidin (Gra) (5 μ M) for 60 min. Cell extracts were collected and mixed with PFO-permeabilized HEK-293T cells expressing ASC and caspase-1 (**a**). After incubation, the reaction mixture was analysed by immunoblotting (**b**). **c**, NLRP3 activity resided in light membrane (P100) and to a lesser extent heavy membrane (P5), but not in cytosol (S100). The fractions were used as ‘activator’ in the NLRP3 activity assay in **a**. **d**, HeLa cells stably expressing NLRP3-GFP were stimulated with 10 μ M nigericin for 80 min before imaging.

¹Department of Molecular Biology, University of Texas Southwestern Medical Center, Dallas, TX, USA. ²Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, TX, USA. *e-mail: zhijian.chen@utsouthwestern.edu

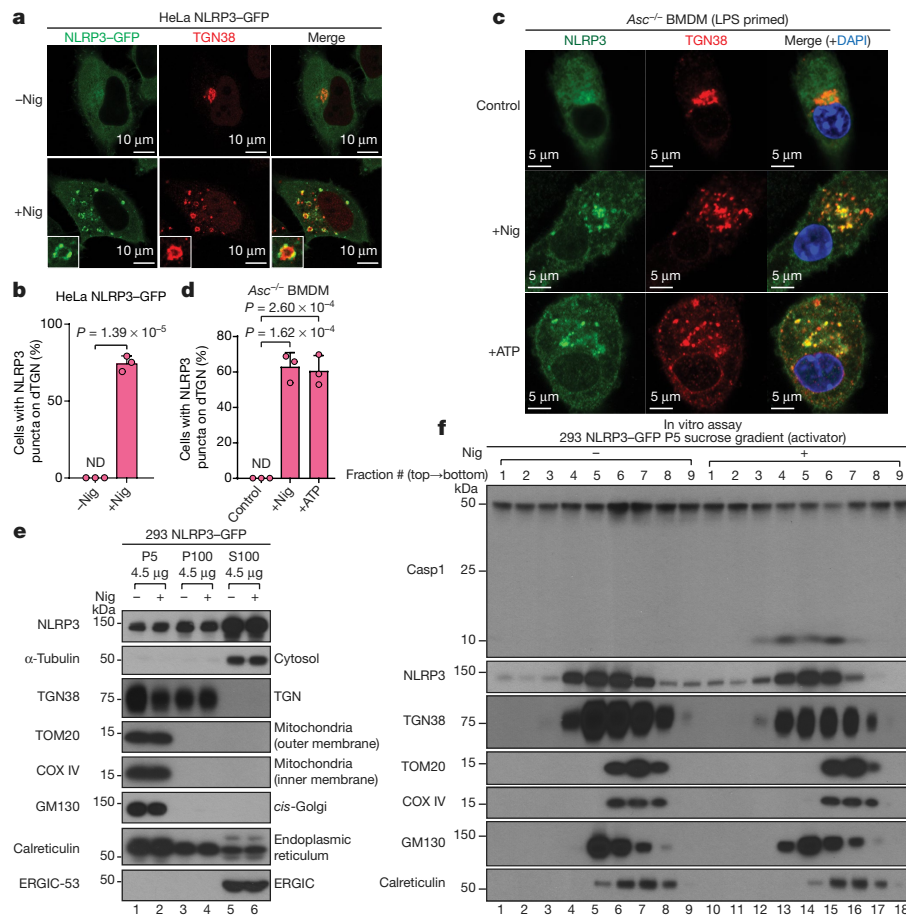


Fig. 2 | NLRP3 aggregates on dTGN. **a, b**, Nigericin induced NLRP3 aggregation on dTGN in HeLa cells stably expressing NLRP3-GFP. The percentage of cells with NLRP3 puncta on dTGN was quantified from 100 cells ($n = 3$, mean \pm s.d., two-sided t -test). ND, not detectable. **c, d**, Nigericin (10 μ M) and ATP (5 mM) both induced endogenous NLRP3 aggregation on dTGN in primary ASC-deficient BMDMs. The cells were primed with 50 ng ml⁻¹ lipopolysaccharide (LPS) for 3 h. The data were

collected and analysed as in **b, e**. Immunoblotting of indicated organelle markers in P5 (heavy membrane), P100 (light membrane) and S100 (cytosol) fractions. ERGIC, endoplasmic reticulum–Golgi intermediate compartment. **f**, The P5 fraction as shown in **e** was further fractionated by sucrose gradient ultracentrifugation followed by NLRP3 activity assay (top panel) or immunoblotting of each fraction (remaining panels).

a bacterial toxin that forms pores in the plasma membrane⁶. After incubation, caspase-1 cleavage was detected in a manner dependent on treatment with NLRP3 stimuli such as nigericin or gramicidin (Fig. 1b). Highly purified NLRP3 also exhibited stimulus-dependent activity in this assay (Extended Data Fig. 1c), suggesting that the activity was intrinsic to NLRP3. The activator cell could also be replaced with RAW 264.7 cells, a macrophage-like cell line that expresses endogenous NLRP3 but not ASC⁷ (Extended Data Fig. 1d). The in vitro assay therefore faithfully recapitulates the activation status of both exogenous and endogenous NLRP3.

Using this in vitro assay, we identified the subcellular fraction in which NLRP3 activity resides. Extracts from HEK-293T cells expressing NLRP3-GFP were fractionated into P5 (heavy membrane), P100 (light membrane) and S100 (cytosol) fractions by differential centrifugation, and the fractions were tested in the NLRP3 activity assay. Although the majority of NLRP3 protein was in S100 (cytosol), NLRP3 activity was only detected in P100 (light membrane) and to a lesser extent P5 (heavy membrane) (Fig. 1c), indicating that only a small fraction of NLRP3 becomes active upon stimulation, and that this fraction is associated with membranes and/or forms large aggregates.

NLRP3 aggregates on dTGN

Fluorescence microscopy experiments showed that NLRP3-GFP was diffused across the cytosol under basal conditions but formed multiple small puncta upon nigericin treatment (Fig. 1d and Extended Data

Fig. 1e). These puncta were distinct from the single large speck formed when ASC was present (Extended Data Fig. 1f), suggesting that NLRP3 first aggregates into multiple small puncta before being incorporated into a large speck with ASC. Enrichment of NLRP3 puncta by saponin treatment showed that these puncta had higher specific activity than NLRP3 in crude cell extracts (Extended Data Fig. 1g), indicating that these puncta are the active form of NLRP3. Notably, the constitutively active disease mutants^{8,9} of NLRP3 could form multiple puncta without stimulation (Extended Data Fig. 1h), again suggesting that NLRP3 requires aggregation to become active.

We observed that NLRP3 puncta were formed on the outside of vesicle-like structures (Fig. 1d and Supplementary Video 1). In cells stimulated with nigericin, a number of giant vesicles appeared in the perinuclear region, typically 0.5–2 μ m in diameter and composed of a single membrane (Extended Data Fig. 2a, b). To understand the origin of these vesicles, we examined various subcellular organelles in HeLa cells stably expressing NLRP3-GFP. Upon stimulation, the *trans* face of the TGN disassembled from a single perinuclear cluster into vesicles on which NLRP3 formed puncta (Fig. 2a, b and Supplementary Video 2).

Whereas the entire TGN disassembled into vesicles (Extended Data Fig. 2c), *cis* and medial Golgi remained intact and did not co-localize with NLRP3 puncta (Extended Data Fig. 2d, e). There was no detectable change in the other organelles examined (see Supplementary Information), with the exception of the early endosomes (Extended Data Fig. 2f), probably because early endosomes continuously exchange

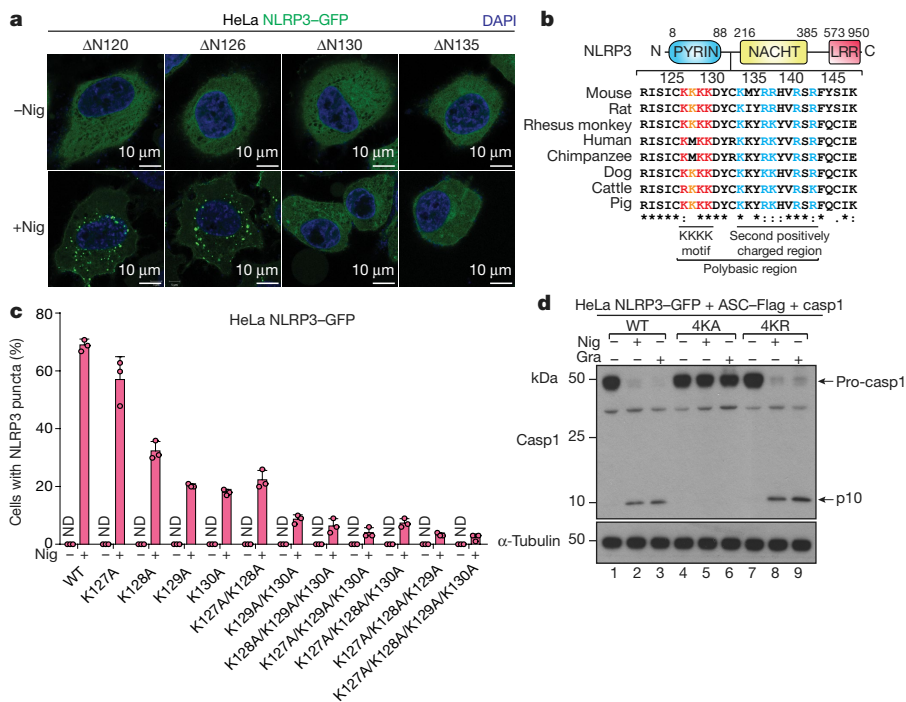


Fig. 3 | NLRP3 is recruited to dTGN via its polybasic region. **a**, HeLa cells stably expressing the N-terminally truncated NLRP3 proteins were stimulated with nigericin. **b**, The polybasic region of NLRP3 is highly conserved in all identified NLRP3 orthologues (aligned using Clustal Omega). LRR, leucine-rich repeat. **c**, Cells stably expressing the indicated

cargo with the TGN. Finally, both ATP and gramicidin, which are structurally unrelated to nigericin, also promoted TGN dispersion and NLRP3 recruitment (Extended Data Fig. 2g, h), although the sizes of dispersed TGN differed depending on the stimulus. By contrast, AIM2

d, HeLa cells stably expressing wild-type NLRP3 (WT), NLRP3(4KA) or NLRP3(4KR) and other indicated proteins were treated with nigericin or gramicidin before immunoblotting.

activation did not involve either TGN disassembly or AIM2 translocation to the TGN (Extended Data Fig. 2i). We refer to the disassembled TGN triggered by NLRP3 stimuli collectively as dTGN.

When primary wild-type bone-marrow-derived macrophages (BMDMs) were stimulated with either nigericin or ATP, substantial TGN disassembly occurred and preceded activation of caspase-1 and IL-1 β (Extended Data Fig. 3a–c). Of note, when ASC-deficient primary BMDMs (used to prevent potential interference by the ASC speck) were treated with nigericin or ATP, endogenous NLRP3 was recruited to the dTGN and formed puncta (Fig. 2c, d). NLRP3 recruitment to dTGN in ASC-deficient BMDMs occurred before the earliest detectable caspase-1 and IL-1 β cleavage in wild-type BMDMs stimulated in the same way (Extended Data Fig. 3b, d, Supplementary Information). Collectively, these results confirm that signal-dependent dTGN formation and endogenous NLRP3 recruitment occur in physiologically relevant cells and precede the activation of the NLRP3 inflammasome.

NLRP3 activity is strongly associated with dTGN

NLRP3 is proposed to be activated by translocation to mitochondria, although this model has been challenged in recent studies^{10,11}. We therefore used imaging and biochemical analyses to further explore the subcellular localization of active NLRP3. We detected no morphological change or co-localization with NLRP3 puncta in mitochondria after stimulation in either reconstituted HeLa cells (Extended Data Fig. 4a and Supplementary Video 3) or primary macrophages (Extended Data Fig. 4b), consistent with these puncta being localized on dTGN (Fig. 2a, c). In an independent approach, we used subcellular fractionation followed by the *in vitro* NLRP3 assay to study the correlation between NLRP3 activity and organelle markers. Figure 1c shows that the majority of NLRP3 activity is present in the P100 fraction, which does not contain mitochondria (Fig. 2e). Indeed, mitochondria were only present in the P5 fraction (Fig. 2e), and did not co-migrate with NLRP3 activity when P5 was further fractionated by sucrose gradient ultracentrifugation (Fig. 2f). By contrast, TGN was present in both P5 and P100 (Fig. 2e), and was the only organelle that strongly co-migrated with NLRP3 activity in both fractions (Fig. 2f and Extended Data

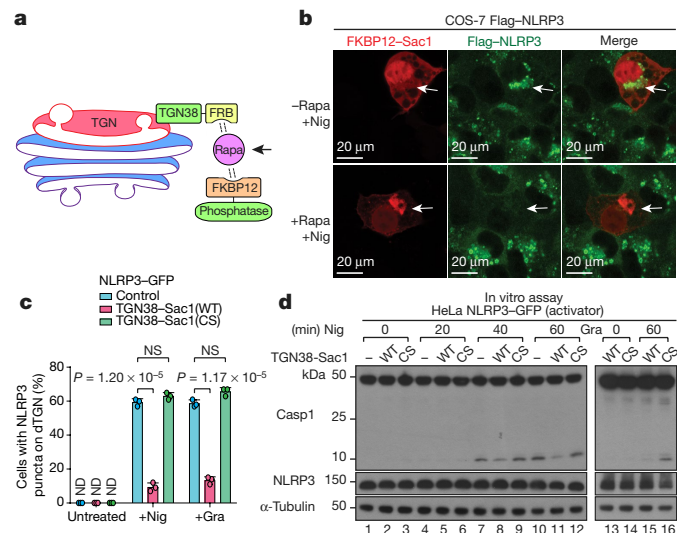


Fig. 4 | NLRP3 is recruited to dTGN via binding to PtdIns4P.

a, Inducible recruitment system of phosphatases: rapamycin (Rapa) (1 μ M) induces the heterodimerization of FRB and FKBP12, thereby recruiting the phosphatase to the TGN to hydrolyze its target phospholipid. **b**, Inducible translocation of the PtdIns4P phosphatase Sac1 to the TGN inhibited nigericin-induced NLRP3 puncta formation in COS-7 cells. Arrows indicate cells with Sac1 expression. **c**, HeLa cells stably expressing the indicated proteins were treated with nigericin or gramicidin. The percentage of cells containing NLRP3 puncta was quantified from 100 cells ($n = 3$, mean \pm s.d., two-sided *t*-test). NS, not significant (significance level, $\alpha = 0.01$). CS, Sac1 (C389S). **d**, Lysates from cells treated as in **c** were analysed using the in vitro NLRP3 activity assay.

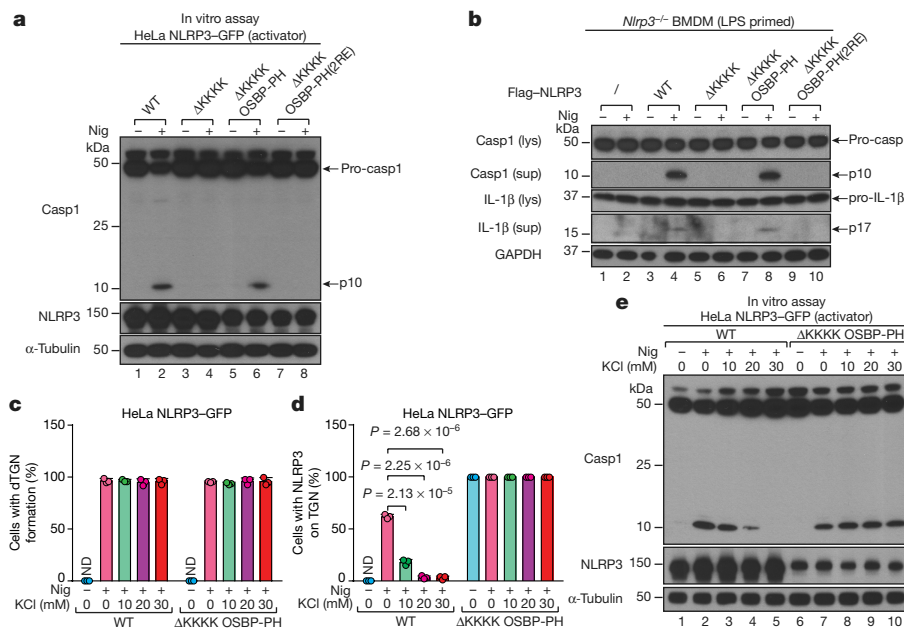


Fig. 5 | NLRP3 binding to PtdIns4P is essential for inflammasome activation. **a**, HeLa cells stably expressing the indicated proteins were treated with nigericin followed by the in vitro NLRP3 activity assay. 2RE, NLRP3(R107/108E). **b**, Primary NLRP3-deficient BMDMs reconstituted with the indicated proteins were primed with LPS before nigericin stimulation. Cell lysates (lys) and supernatants (sup) were analysed by immunoblotting to detect cleaved caspase-1 and IL-1 β . / indicates

no rescue by an NLRP3 expression vector. **c**, Cells were treated with nigericin in the presence of increasing concentrations of KCl, and the percentage of cells with dTGN formation was quantified from 100 cells ($n = 3$, mean \pm s.d., two-sided t -test). **d**, Cells were treated as in **c**, and the percentage of cells with NLRP3 on the TGN (either intact TGN or dTGN) was analysed as in **c**. **e**, Cells were treated as in **c**; cell lysates were analysed by the in vitro NLRP3 activity assay.

Fig. 4c). Finally, oligomerization of the pyrin domain of ASC (ASC-PYD, residues 1–90 of ASC) was initiated from dTGN-localized NLRP3 puncta (Extended Data Fig. 4d, Supplementary Information). Together, these data strongly indicate that NLRP3 becomes active following recruitment to dTGN.

NLRP3 is recruited to dTGN via its polybasic region

dTGN formation occurred earlier than NLRP3 puncta formation (Supplementary Video 2), and was independent of NLRP3 (Extended Data Fig. 5a, b and Supplementary Video 4), suggesting that the presence of dTGN may be a prerequisite for the recruitment, aggregation and activation of NLRP3. This is consistent with the observation that the constitutively active NLRP3 disease mutants could bypass TGN recruitment because they were able to aggregate without stimulation (Extended Data Fig. 5c).

To investigate how NLRP3 is recruited to dTGN, we examined several N-terminally truncated mutants of NLRP3 and found that deletion of four consecutive lysine residues (residues 127 to 130, hereafter referred to as the KKKK motif) abruptly abolished NLRP3 aggregation (Fig. 3a). The KKKK motif is a highly conserved region between the pyrin domain and the NACHT domain, which contains at least three positively charged residues in all known NLRP3 orthologues (Fig. 3b). Mutations of residues in the KKKK motif to alanine inhibited the ability of full-length NLRP3 to form puncta in a manner dependent on the number of remaining lysine residues, with the K127A/K128A/K129A/K130A (4KA) mutant forming almost no detectable puncta (Fig. 3c and Extended Data Fig. 5d). All NLRP3 mutants that were compromised in puncta formation also exhibited defective activation of downstream signalling (Extended Data Fig. 5e). The 4KA mutation did not affect the ability of the constitutively active mutant L351P to aggregate in the cytosol or activate downstream signalling without stimulation (Extended Data Fig. 5f), indicating that the 4KA mutation did not block the interaction of NLRP3 with ASC directly.

When the lysines in the KKKK motif were mutated to arginine (4KR), puncta formation and activation of NLRP3 by nigericin were similar

to those of the wild-type protein (Extended Data Fig. 6a, b), indicating that the positive charge of the motif is critical for NLRP3 recruitment and activation. Similarly, NLRP3(4KA) but not NLRP3(4KR) was largely defective in puncta formation in response to gramicidin or ATP (Extended Data Fig. 6c). Finally, NLRP3(4KR), but not NLRP3(4KA), retained the ability to induce ASC oligomerization (Extended Data Fig. 6d) and caspase-1 cleavage (Fig. 3d and Extended Data Fig. 6e). We also identified a second positively charged region located after the KKKK motif (Fig. 3b), which was also important for the recruitment and activation of NLRP3 (Extended Data Fig. 6f, g, Supplementary Information). Together, these results demonstrate that the conserved polybasic region including the KKKK motif has a critical function in recruiting NLRP3 to dTGN upon stimulation, which is essential for its subsequent activation.

NLRP3 is recruited to dTGN via binding to PtdIns4P

Several Rab GTPases are known to bind to negatively charged phospholipids PtdIns(3,4,5)P₃ and PtdIns(4,5)P₂ on the plasma membrane via polybasic regions¹². To test whether the polybasic region in NLRP3 mediates its recruitment through a similar mechanism, we purified a fragment of NLRP3 containing the polybasic region (residues 127–146) and tested its binding to phospholipids by lipid blot assay. This fragment bound to several negatively charged phospholipids, and the binding was abolished by the 4KA mutation (Extended Data Fig. 7a). Next, we investigated which phospholipid is important for recruitment of full-length NLRP3 to dTGN in live cells. We took advantage of the inducible recruitment system of phospholipid phosphatases¹³, in which the addition of rapamycin promotes the heterodimerization of FKBP12 and FRB, thereby recruiting the phosphatase to the TGN, where it hydrolyzes its target phospholipid (Fig. 4a). Sac1, a PtdIns4P phosphatase, abolished NLRP3 recruitment to dTGN in a manner dependent on the translocation of Sac1 to the TGN (Fig. 4b) and its catalytic activity (Extended Data Fig. 7b). Sac2¹⁴, another PtdIns4P phosphatase, also abolished NLRP3 recruitment, whereas phosphatases targeting other phospholipids had no detectable effect on NLRP3 recruitment (Extended Data Fig. 7c).

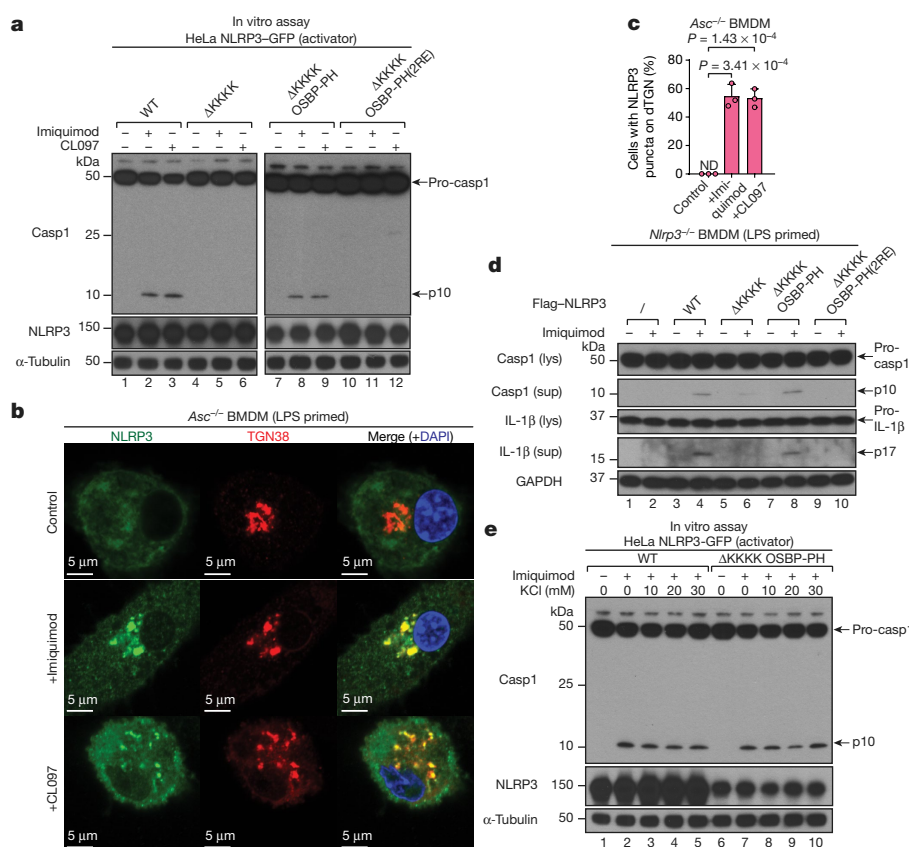


Fig. 6 | Binding to PtdIns4P on dTGN is essential for K^+ -efflux-independent NLRP3 activation. **a**, HeLa cells stably expressing the indicated proteins were treated with $45 \mu\text{g ml}^{-1}$ imiquimod or CL097 for 80 min before cell lysates were analysed by the in vitro NLRP3 activity assay. **b**, **c**, Primary ASC-deficient BMDMs were primed with LPS and incubated with $45 \mu\text{g ml}^{-1}$ imiquimod or CL097 for 60 min. The percentage of cells with NLRP3 puncta on the dTGN was quantified from 100 cells ($n = 3$, mean \pm s.d., two-sided t -test). **d**, Primary NLRP3-

deficient BMDMs reconstituted with the indicated proteins were primed with LPS and treated with imiquimod. Cell lysates and supernatants were analysed by immunoblotting with the indicated antibodies to assess activation of the inflammasome. / indicates no rescue by an NLRP3 expression vector. Lys, lysate; sup, supernatant. **e**, HeLa cells stably expressing the indicated NLRP3 proteins were treated with imiquimod in the presence of increasing concentrations of KCl and the cell lysates were analysed with the in vitro NLRP3 activation assay.

To better quantify the impact of PtdIns4P on NLRP3 recruitment and activation, we stably expressed TGN38–Sac1 fusion protein in HeLa cells also expressing NLRP3–GFP (Extended Data Fig. 7d). Expression of TGN38–Sac1 did not affect the general cellular morphology or nigericin-induced dTGN formation (Extended Data Fig. 7e), but largely impaired NLRP3 puncta formation through its phosphatase activity (Fig. 4c). TGN38–Sac1 did not affect poly(dA:dT)-induced aggregation of AIM2 (Extended Data Fig. 7f). Moreover, nigericin-induced NLRP3 puncta strongly co-localized with the PH domain of OSBP (OSBP-PH), one of the best characterized PtdIns4P-binding domains¹⁵, but not with the AP-1 complex (Extended Data Fig. 7g), which relies mainly on ARF1 for TGN targeting^{16,17}, suggesting that NLRP3 is specifically recruited to PtdIns4P-enriched microdomains on the TGN. Finally, TGN38–Sac1 markedly repressed NLRP3 activation through its catalytic activity (Fig. 4d). Together, these results show that PtdIns4P binding is important for NLRP3 recruitment to the dTGN and its subsequent activation, which is consistent with the observation that PtdIns4P is enriched on the TGN¹⁸.

As expected, deletion of the KKKK motif (NLRP3(Δ KKKK)) abolished the recruitment of NLRP3 to the dTGN, whereas replacement of the KKKK motif with OSBP-PH (NLRP3(Δ KKKK/OSBP-PH)) constitutively targeted NLRP3 to the TGN (Extended Data Fig. 8a, b). By contrast, insertion of OSBP-PH(R107/108E), which harbours mutations that abolish its binding to PtdIns4P¹⁹, did not enable NLRP3(Δ KKKK/OSBP-PH(R107/108E)) to translocate to the TGN (Extended Data Fig. 8a). Notably, NLRP3(Δ KKKK) exhibited little activity after stimulation, which could be rescued by substituting it with NLRP3(Δ KKKK/OSBP-PH), but not by NLRP3

(Δ KKKK/OSBP-PH(R107/108E)), in both reconstituted HeLa cells and primary NLRP3-deficient BMDMs (Fig. 5a, b and Extended Data Fig. 8c, d). Targeting to PtdIns4P-enriched microdomains, and not general TGN localization, is essential for NLRP3 activation (Extended Data Fig. 8e, Supplementary Information). These results confirm that the KKKK motif serves as a PtdIns4P-binding domain for NLRP3 recruitment and activation.

To test whether potassium (K^+) efflux, a cellular signal that has been shown to be important for NLRP3 activation²⁰, is required for dTGN formation, NLRP3 recruitment and/or TGN-localized NLRP3 to become active, we investigated the sensitivity of wild-type NLRP3 and NLRP3(Δ KKKK/OSBP-PH) to increasing concentrations of KCl in the extracellular medium, which abolished stimulus-induced K^+ efflux²⁰ (Extended Data Fig. 8f). Inhibition of K^+ efflux had no effect on dTGN formation (Fig. 5c). Surprisingly, whereas inhibition of K^+ efflux repressed the recruitment of wild-type NLRP3 to dTGN and blocked its activation, it had no effect on either constitutive TGN localization or signal-dependent activation of NLRP3(Δ KKKK/OSBP-PH) (Fig. 5d, e). This indicates that K^+ efflux is essential for recruitment of wild-type NLRP3 to the TGN, probably by lowering cellular ionic strength to promote ionic binding; however, once NLRP3 is recruited to the TGN, such as in the case of NLRP3(Δ KKKK/OSBP-PH), it no longer requires K^+ efflux for subsequent activation. Of note, even though NLRP3(Δ KKKK/OSBP-PH) was targeted to TGN constitutively and no longer required K^+ efflux, its activation was still signal-dependent. This is likely to indicate that binding to PtdIns4P on dispersed TGN rather than intact TGN is essential for NLRP3 to become active. This is consistent with the observation that spontaneous K^+ efflux

alone is not sufficient for NLRP3 activation in both reconstituted HeLa cells (Extended Data Fig. 8g, h) and primary BMDMs (Extended Data Fig. 8i, j), because K^+ efflux does not induce TGN dispersion.

dTGN enables K^+ -efflux-independent NLRP3 activation

Recent studies have shown that NLRP3 can be activated by the immune modulators imiquimod and CL097 in a K^+ -efflux-independent manner^{21,22}. We found that both imiquimod and CL097 also induced marked dispersion of TGN and recruitment of NLRP3 to dTGN. Moreover, formation of NLRP3 puncta was blocked by deleting the KKKK motif, and could be restored by replacing the KKKK motif with OSBP-PH but not OSBP-PH(R107/108E) (Extended Data Fig. 9a, b). Furthermore, both imiquimod- and CL097-mediated NLRP3 activation was abolished when the KKKK motif of NLRP3 was deleted, and could be rescued by the insertion of OSBP-PH but not OSBP-PH(R107/108E) (Fig. 6a). Similarly, endogenous NLRP3 was recruited to dTGN (Fig. 6b, c) but not mitochondria (Extended Data Fig. 9c) after treatment with imiquimod or CL097 in primary ASC-deficient BMDMs. Moreover, the KKKK motif was essential for NLRP3 to restore caspase-1 and IL-1 β cleavage in primary NLRP3-deficient BMDMs in response to imiquimod, and the KKKK motif could be replaced by OSBP-PH but not by OSBP-PH(R107/108E) (Fig. 6d and Extended Data Fig. 8d). Therefore, PtdIns4P-mediated recruitment to dTGN is also essential for K^+ -efflux-independent NLRP3 activation.

Consistent with the K^+ -efflux-independence of imiquimod and CL097 as stimuli²², the recruitment of NLRP3 to dTGN (Extended Data Fig. 9d) and the activation of both wild-type NLRP3 and NLRP3(Δ KKKK/OSBP-PH) (Fig. 6e) were highly resistant to extracellular KCl. It is unclear what contributes to the differences in the requirement of K^+ efflux for different NLRP3 stimuli. One possibility is that K^+ -efflux-independent stimuli induced much more pronounced dispersion of TGN and partial separation of PtdIns4P from other TGN compartments such as those marked by TGN38 (Extended Data Fig. 9a, Supplementary Information), which may help to expose PtdIns4P in a more efficient conformation to recruit NLRP3 even without the induction of K^+ efflux.

In summary, using complementary techniques in both reconstituted systems and primary macrophages, we have identified a common cellular signal triggered by diverse NLRP3 stimuli: the disassembly of TGN into dTGN, to which NLRP3 is recruited through ionic bonding between its polybasic region and PtdIns4P on dTGN (Extended Data Fig. 10). dTGN then serves as a scaffold for NLRP3 to aggregate and interact with ASC, which activates the downstream signalling cascade. This mechanism of NLRP3 activation is reminiscent of the 'guard model' in plants, in which the disease resistant (R) proteins indirectly recognize virulent factors by monitoring the integrity of host targets, the 'pathogen-induced altered self'^{23,24}. By binding to dTGN as the 'altered self', NLRP3 indirectly senses a large variety of molecules associated with pathogens and other dangers.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0761-3>.

Received: 28 May 2017; Accepted: 17 October 2018;
Published online 28 November 2018.

1. Broz, P. & Dixit, V. M. Inflammasomes: mechanism of assembly, regulation and signalling. *Nat. Rev. Immunol.* **16**, 407–420 (2016).
2. Lamkanfi, M. & Dixit, V. M. Inflammasomes and their roles in health and disease. *Annu. Rev. Cell Dev. Biol.* **28**, 137–161 (2012).
3. Bryan, N. B., Dorfleutner, A., Rojanasakul, Y. & Stehlik, C. Activation of inflammasomes requires intracellular redistribution of the apoptotic speck-like protein containing a caspase recruitment domain. *J. Immunol.* **182**, 3173–3182 (2009).

4. Lu, A. et al. Unified polymerization mechanism for the assembly of ASC-dependent inflammasomes. *Cell* **156**, 1193–1206 (2014).
5. Cai, X. et al. Prion-like polymerization underlies signal transduction in antiviral immune defense and inflammasome activation. *Cell* **156**, 1207–1222 (2014).
6. Rossjohn, J. et al. Structures of perfringolysin O suggest a pathway for activation of cholesterol-dependent cytolysins. *J. Mol. Biol.* **367**, 1227–1236 (2007).
7. Gross, O. Measuring the inflammasome. *Methods Mol. Biol.* **844**, 199–222 (2012).
8. Meng, G., Zhang, F., Fuss, I., Kitani, A. & Strober, W. A mutation in the *Nlrp3* gene causing inflammasome hyperactivation potentiates Th17 cell-dominant immune responses. *Immunity* **30**, 860–874 (2009).
9. Brydges, S. D. et al. Inflammasome-mediated disease animal models reveal roles for innate but not adaptive immunity. *Immunity* **30**, 875–887 (2009).
10. Juliana, C. et al. Non-transcriptional priming and deubiquitination regulate NLRP3 inflammasome activation. *J. Biol. Chem.* **287**, 36617–36622 (2012).
11. Bauernfeind, F. et al. Cutting edge: reactive oxygen species inhibitors block priming, but not activation, of the NLRP3 inflammasome. *J. Immunol.* **187**, 613–617 (2011).
12. Heo, W. D. et al. PI(3,4,5)P₃ and PI(4,5)P₂ lipids target proteins with polybasic clusters to the plasma membrane. *Science* **314**, 1458–1461 (2006).
13. Szentpetery, Z., Varnai, P. & Balla, T. Acute manipulation of Golgi phosphoinositides to assess their importance in cellular trafficking and signaling. *Proc. Natl Acad. Sci. USA* **107**, 8225–8230 (2010).
14. Hsu, F., Hu, F. & Mao, Y. Spatiotemporal control of phosphatidylinositol 4-phosphate by Sac2 regulates endocytic recycling. *J. Cell Biol.* **209**, 97–110 (2015).
15. Balla, T. & Varnai, P. Visualizing cellular phosphoinositide pools with GFP-fused protein-modules. *Sci. STKE* **2002**, pl3 (2002).
16. Traub, L. M., Ostrom, J. A. & Kornfeld, S. Biochemical dissection of AP-1 recruitment onto Golgi membranes. *J. Cell Biol.* **123**, 561–573 (1993).
17. Boman, A. L., Zhang, C., Zhu, X. & Kahn, R. A family of ADP-ribosylation factor effectors that can alter membrane transport through the trans-Golgi. *Mol. Biol. Cell* **11**, 1241–1255 (2000).
18. Di Paolo, G. & De Camilli, P. Phosphoinositides in cell regulation and membrane dynamics. *Nature* **443**, 651–657 (2006).
19. Levine, T. P. & Munro, S. Targeting of Golgi-specific pleckstrin homology domains involves both PtdIns 4-kinase-dependent and -independent components. *Curr. Biol.* **12**, 695–704 (2002).
20. Munoz-Planillo, R. et al. K^+ efflux is the common trigger of NLRP3 inflammasome activation by bacterial toxins and particulate matter. *Immunity* **38**, 1142–1153 (2013).
21. Kanneganti, T. D. et al. Bacterial RNA and small antiviral compounds activate caspase-1 through cryopyrin/Nalp3. *Nature* **440**, 233–236 (2006).
22. Gross, C. J. et al. K^+ efflux-independent NLRP3 inflammasome activation by small molecules targeting mitochondria. *Immunity* **45**, 761–773 (2016).
23. Bonardi, V., Cherkis, K., Nishimura, M. T. & Dangi, J. L. A new eye on NLR proteins: focused on clarity or diffused by complexity? *Curr. Opin. Immunol.* **24**, 41–50 (2012).
24. Khan, M., Subramaniam, R. & Desveaux, D. Of guards, decoys, baits and traps: pathogen perception in plants by type III effector sensors. *Curr. Opin. Microbiol.* **29**, 49–55 (2016).

Acknowledgements We thank T. Balla and M. Seaman for reagents, C. Pasare for sharing bones from ASC-deficient mice, B. Beutler for NLRP3-deficient mice, Molecular Biology Imaging Facility, Electron Microscopy Core Facility and O'Brien Kidney Research Center at UT Southwestern Medical Center for technical assistance, and all members of the Chen laboratory for their help and support. This work was supported by grants from the Welch Foundation (I-1389) and the Cancer Prevention and Research Institute of Texas (RP150498 and RP110430). Z.J.C. is a Howard Hughes Medical Institute Investigator.

Reviewer information Nature thanks J. Vince and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.C. designed the study under the guidance of Z.J.C., performed all the experiments, analysed the data and prepared the manuscript. Z.J.C. designed and supervised this study and revised the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0761-3>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0761-3>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to Z.J.C.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Antibodies and chemicals. Antibodies against NLRP3 (AG-20B-0014) and ASC (AG-25B-0006) were from AdipoGen. Antibodies against caspase-1 (sc-515), human TGN38 (sc-33783) and TOM20 (sc-11415) were from Santa Cruz Biotechnology. Antibody against murine TGN38 was a gift from M. Seaman (Cambridge Institute for Medical Research). Antibodies against GOLGA4 (611280) and GM130 (610822) were from BD Biosciences. Antibodies against tubulin (T5168), flag (F1804) and AP1G1 (A4200) were from Sigma. Antibody against haemagglutinin (HA) (mms-101p) was from Covance. Antibody against giantin (ab24586) was from Abcam. Antibody against COX IV (20E8C12) was from Invitrogen. Antibodies against calreticulin (2891) and GAPDH (2118) were from Cell Signaling. Antibody against IL-1 β (AB-401-NA) was from R&D Systems. Antibody against ERGIC-53 (ALX-804-602-C100) was from Axxora. Alexa Fluor secondary antibodies (488, 568, and 633) were from Life Technologies.

Flag antibody-conjugated agarose (A2220), nigericin (N7143) and ATP (A7699) were from Sigma. Gramicidin (ALX-350-233) was from Enzo Life Sciences. Rapamycin (AG-CN2-0025) was from AdipoGen. LPS (Ultra-Pure), imiquimod (tlrl-imq) and CL097 (tlrl-c97) were from InvivoGen.

Mammalian cell culture. Cell lines, including HEK-293T, HeLa, COS-7 and RAW 264.7 were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% (v/v) cosmic calf serum (Hyclone), penicillin (100 U/ml) and streptomycin (100 μ g/ml). These cell lines were originally obtained from ATCC (<https://www.atcc.org/>), and were regularly monitored for contamination from other cell lines. They were free of mycoplasma contamination, based on the results of e-Myco Mycoplasma PCR Detection Kit (Bulldog Bio), and were regularly maintained with Normocin (an antimicrobial reagent against mycoplasma, bacteria and fungi) (InvivoGen).

For primary BMDM induction, cells were isolated from mouse bone marrow and cultured in mCSF-1-containing RPMI 1640 medium supplemented with 10% (v/v) fetal bovine serum (Hyclone), penicillin and streptomycin as mentioned above. All cells were cultured at 37°C in an atmosphere with 5% (v/v) CO₂. NLRP3-deficient mice (B6.129S6-Nlrp3^{tm1Bhk/J}) were provided by B. Beutler (UT Southwestern Medical Center), and were originally from The Jackson Laboratory (Stock No: 021302). Bones from ASC-deficient mice were provided by C. Pasare (UT Southwestern Medical Center). All mice were bred and maintained under specific pathogen-free conditions in the animal care facility of University of Texas Southwestern Medical Center according to experimental protocols approved by the Institutional Animal Care and Use Committee.

Inflammasome stimulation. For stimulation of reconstituted cell lines, the culture medium was replaced by OPTI-MEM medium (Life Technologies) containing inflammasome stimulus, including nigericin (10 μ M), gramicidin (5 μ M), ATP (5 mM, pH adjusted to 7.5), imiquimod (45 μ g/ml) or CL097 (45 μ g/ml). The cells were then incubated at 37°C in an atmosphere with 5% (v/v) CO₂ for 60 min (HEK-293T) or 80 min (HeLa, COS-7 and BJ) unless otherwise specified. Poly(dA:dT) (Sigma) was transfected into cells with Lipofectamine 2000 (Invitrogen) at 1.5 μ g/ml for 3 h. Priming with TLR ligands is not required for reconstituted cell lines owing to the high expression level of stably introduced NLRP3, in line with previous studies^{10,25}.

For stimulation of primary BMDMs or RAW 264.7 cells, cells were primed with LPS (50 ng/ml) for 3 h before NLRP3 stimulus was added: nigericin (10 μ M), ATP (5 mM, pH adjusted to 7.5), imiquimod (45 μ g/ml) or CL097 (45 μ g/ml) for 60 min unless otherwise specified.

Cells were collected in lysis buffer A (20 mM Tris-HCl (pH 7.5), 150 mM NaCl, 0.5% NP-40, 1 mM DTT, protease inhibitor cocktail (Roche)) and centrifuged at 20,000g for 15 min to collect lysate for immunoblotting. For primary BMDMs, the medium (supernatant) was also collected for detection of secreted caspase-1 p10 and IL-1 β p17.

Plasmids, viruses and stable cell lines. The lentiviral vectors for expressing proteins, pTY-EF1a-puroR/hygroR/zeoR-2A-GFP-Flag, were modified from a vector kindly provided by Y. Zhang (University of North Carolina at Chapel Hill). These vectors were further modified into pTY-EF1a-GFP-IRES-puroR/hygroR/zeoR, to circumvent the problem of incomplete cleavage by the 2A protease.

For protein expression, cDNA encoding proteins were purchased from Life Technologies or Open Biosystems and cloned into the lentiviral vectors described above. NLRP3, ASC and caspase-1 genes were of mouse origin, with the exception of the human NLRP3 used in Extended Data Fig. 1e. For NLRP3(4KA)-GFP-GOLGA4^{GRIP}, the C-terminal 312 amino acids of the mouse *Golga4* gene were fused to the C terminus of coding sequence for NLRP3(K127A/K128A/K129A/K130A)-GFP separated by a short linker (RSIAT). All the other genes used in this study are of human origin unless otherwise specified. HomoloGene database (NCBI) was used to search for all currently identified NLRP3 orthologues and the results were further confirmed by Protein BLAST.

OSBP-PH-GFP consists of an initial methionine residue and the PH domain from OSBP (human, amino acids 87–185) followed by a short linker (RSIAT) and

GFP. For NLRP3(Δ KKKK), NLRP3 (Δ KKKK/OSBP-PH) and NLRP3(Δ KKKK/OSBP-PH(R107/108E)), a short linker (GGGGS) was inserted at the position of deleted amino acids 127–130 to maintain the structural flexibility between different domains. ASC-PYD consists of only the pyrin domain of mouse ASC (amino acids 1–90).

MitoDsRed2 protein was stably expressed in HeLa cells expressing NLRP3-GFP by infection with lentiviral vector pCDH-DsRed2-Mito. The DsRed2-Mito gene in this lentiviral vector was cloned from pDsRed2-Mito (Clontech, 632421), which consists of DsRed2 fused to the mitochondrial targeting sequence of COX VIII at its N terminus. The plasmid containing the coding sequence of PFO from the bacterium *Clostridium perfringens* was provided by R. DeBose-Boyd (UT Southwestern Medical Center).

With the exception of the inducible phosphatase recruitment system, all proteins were stably expressed through lentiviral infection, which was performed as described previously²⁶. In brief, lentivirus was packaged by co-transfecting HEK-293T cells with the lentiviral vector and packaging vectors. Medium containing lentivirus was filtered and added to target cells in the presence of polybrene (10 μ g/ml). Cells were then selected with the respective antibiotics for at least seven days before protein expression was confirmed by immunoblotting and fluorescence microscopy.

In vitro assay for NLRP3 activation. As shown in the schematic in Fig. 1a, two stable HEK-293T cell lines were set up that stably expressed only NLRP3 (293 NLRP3) or only ASC and caspase-1 (293 ASC-casp1). 293 NLRP3 cells were incubated with or without inflammasome stimulus (for example, nigericin), before the cells were collected in lysis buffer B (10 mM Tris-HCl (pH 7.5), 10 mM KCl, 1.5 mM MgCl₂, 0.2% NP-40, protease inhibitor cocktail) and centrifuged at 1,000g for 5 min. Supernatant (cell extracts) (10 μ g) of 293 NLRP3 ('activator') was then mixed with 10⁶ cells from 293 ASC-casp1 ('recipient') semi-permeabilized with 60 ng PFO and additional lysis buffer B was added to reach a final volume of 10 μ l. After incubation at 30°C for 80 min, the reaction mixture was incubated with additional 0.5% NP-40 on ice for 15 min before centrifugation at 20,000g for 15 min. The supernatant was boiled in SDS loading buffer and used for caspase-1 immunoblotting. In later experiments, other cells that expressed NLRP3 but not ASC (for example, HeLa NLRP3-GFP and RAW 264.7) were also used as activator cells. In addition, ASC and caspase-1 in recipient cells can be replaced by the fusion protein ASC-PYD-p20-p10 consisting of the pyrin domain (PYD, amino acids 1–104) from ASC and the C-terminal p20-p10 region from caspase-1 (amino acids 92–402), a fusion protein reported to bypass the CARD-CARD interaction between ASC and caspase-1²⁷.

Fractionation of HEK-293Ts expressing NLRP3-GFP and purification of NLRP3. Twenty plates (15 cm in diameter) of HEK-293T cells stably expressing Flag-NLRP3-GFP were incubated with or without nigericin, homogenized in isotonic buffer (0.25 M sucrose, 10 mM Tris-HCl (pH 7.5), 10 mM KCl, 1.5 mM MgCl₂ and protease inhibitor cocktail) and centrifuged at 1,000g for 5 min to remove nucleus pellet (P1). The supernatant (S1) was further centrifuged at 5,000g for 10 min to obtain heavy membrane fraction (pellet, P5), while this supernatant (S5) was centrifuged at 100,000g for 20 min to separate light membrane fraction (pellet, P100) from cytosol fraction (supernatant, S100). P5 and P100 were washed with isotonic buffer once and resuspended in the same buffer.

P5 and P100 fractions were then used for sucrose gradient ultracentrifugation separately. In brief, sucrose solutions in low salt buffer (10 mM Tris-HCl (pH 7.5), 10 mM KCl, 1.5 mM MgCl₂) were loaded in a centrifuge tube (from bottom to top, 60%, 50%, 40%, 30%, 20%, 2.1 ml per layer). P5 or P100 (0.8 mg) was loaded on top of the gradient and centrifuged at 170,000g for 2 h. Nine fractions (1.2 ml per fraction) were collected from top to bottom. For each fraction, 3 μ l (for P5) or 1 μ l (for P100) was used for the in vitro activity assay and 10 μ l was used for immunoblotting against various proteins.

Fraction number 4 (from top) of P5 sucrose gradient was used for further purification. Fraction number 4 was first incubated with 0.2% NP-40 on ice for 20 min before centrifugation at 5,000g for 10 min. The supernatant (fraction extract) was immunoprecipitated with Flag M2 agarose in the presence of additional KAc (110 mM) and NaCl (100 mM) at 4°C for 6 h. The agarose beads were then washed five times with Flag IP wash buffer A (10 mM Tris-HCl (pH 7.5), 10 mM KCl, 1.5 mM MgCl₂, 110 mM KAc, 100 mM NaCl, 0.05% NP-40) before elution in low salt buffer with Flag peptides overnight. The eluate was filtered and concentrated using a 10-kDa cut-off concentrator before the purity of NLRP3 was confirmed by both silver staining and mass spectrometry.

Enrichment of nigericin-induced NLRP3 puncta. HEK-293T cells expressing NLRP3-GFP cells were incubated with or without nigericin for 60 min, before washing with PBS twice and incubating with regular PBS or PBS containing 0.3% saponin for 10 min. The cells (still attached to plates) were then washed with PBS again and imaged by fluorescence microscopy. The cells were then collected and lysed with lysis buffer B, before the cell extracts were used as 'activator' for the in vitro assay.

Functional rescue of primary NLRP3-deficient BMDMs. Bone marrow cells were collected from NLRP3-deficient mice and cultured in mCSF-1-containing medium. On day 2 and day 3, respectively, cells were infected with lentivirus encoding pTY-EF1a-zeoR-2A-Flag-NLRP3 (wild type or mutants). As a control, an equal volume of medium without lentivirus was added to cells. On day 7 post-induction in mCSF-1-containing medium, cells were primed with LPS (50 ng/ml) for 3 h (to induce pro-IL-1 β) before nigericin (10 μ M) or imiquimod (45 μ g/ml) was added for another 60 min. The cell lysate and medium (supernatant) were then collected for immunoblotting.

K⁺ efflux inhibition and induction. To inhibit nigericin-induced K⁺ efflux, additional KCl was added into OPTI-MEM medium at the indicated final concentrations at the same time with nigericin. Up to 30 mM of extracellular KCl has been reported to specifically inhibit NLRP3 inflammasome but not other inflammasome pathways or general cell signalling²².

To induce spontaneous K⁺ efflux without adding NLRP3 stimuli, cells were incubated in K⁺-free Hanks' buffer (145 mM NaCl, 1.3 mM CaCl₂, 1.0 mM MgSO₄, 10 mM HEPES (pH 7.5), 5.5 mM glucose) for the indicated time period. As a control, cells were incubated in regular Hanks' buffer (140 mM NaCl, 5 mM KCl, 1.3 mM CaCl₂, 1.0 mM MgSO₄, 10 mM HEPES (pH 7.5), 5.5 mM glucose) with or without nigericin (10 μ M) for a similar time period.

Measurement of intracellular K⁺ concentration. Cells were washed with ice-cold PBS and centrifuged at 1,000g for 5 min to remove residual buffer; this process was repeated once. The cell pellets were then resuspended in Milli-Q H₂O at a volume equal to the cell pellet volume, before being snap-frozen with liquid nitrogen and thawed repeatedly for a total of six cycles to lyse the cells. The lysed cells were centrifuged at 14,000g for 10 min and the cell extracts (supernatant) were collected. The volumes of cell extracts were further adjusted according to protein concentrations, before measured by Jenway Flame Photometer (FPF7) for K⁺ concentration, with help from M. Baum and S. Legan at O'Brien Kidney Research Center (UT Southwestern Medical Center). A standard curve was made for each experiment with titrated KCl solutions to ensure the results fell within the linear range of measurement.

Immunostaining and fluorescence microscopy. For immunostaining, cells were fixed with 4% paraformaldehyde and permeabilized with 0.1% Triton X-100 in PBS, before incubation with primary antibodies followed by Alexa Fluor secondary antibodies. Nuclei were stained with DAPI in mounting medium (Vectashield). Subsequently, we also used 0.1% saponin in place of 0.1% Triton X-100 in the permeabilization step to better preserve the dTGN structures in fixed cells. For immunostaining of endogenous NLRP3 in primary BMDMs, Tyramide Signal Amplification Kits (T20948 and T20949, Life Technologies) were used to enhance the signal-to-noise ratio. For imaging with multiple channels, extensive controls were performed to make sure there was no non-specific staining or crosstalk between channels. These controls include: a) using cells that lack one of the proteins of interest; and/or b) performing staining without one of the primary or secondary antibodies.

Fluorescence images of fixed cells were taken with a Zeiss LSM 700 confocal laser scanning microscope. Time-lapse imaging of live cells was performed using a Nikon A1R microscope equipped with Tokai Hit Incubator System. Phase contrast and fluorescence imaging of live cells were taken with an EVOS FL Cell Imaging System (Thermo Fisher Scientific).

Transmission electron microscopy. HeLa cells were incubated with or without nigericin for 80 min before being fixed with 2.5% (v/v) glutaraldehyde in 0.1 M sodium cacodylate buffer. After rinsing with 0.1 M sodium cacodylate buffer three times, the cells were post-fixed in 1% osmium tetroxide and 0.8% potassium ferricyanide in 0.1 M sodium cacodylate buffer for 1 h. Cells were then rinsed with water and stained with 2% aqueous uranyl acetate overnight. After rinsing with water three times, cells were dehydrated with increasing concentrations of ethanol, infiltrated with Embed-812 resin and polymerized in a 60 °C oven overnight. Blocks were sectioned with a diamond knife (Diatome) on a Leica Ultracut UC7 ultramicrotome (Leica Microsystems) and collected onto copper grids, before post-staining with 2% uranyl acetate in water and lead citrate. Images were acquired on a Tecnai G2 spirit transmission electron microscope (FEI) equipped with a LaB6 source using a voltage of 120 kV with the help from Electron Microscopy Core Facility (UT Southwestern Medical Center).

PIP Strip assay. The following proteins were purified from HEK-293T stable cell lines: Flag-GFP, Flag-NLRP3(127–146)-GFP, and Flag-NLRP3(127–146, K127A/K128A/K129A/K130A)-GFP. The C-terminal GFP tag was introduced to increase the stability of short fragments expressed in mammalian cells. These cells were lysed in lysis buffer C (10 mM Tris-HCl (pH 7.5), 10 mM KCl, 1.5 mM MgCl₂, 500 mM NaCl, 0.2% NP-40, and protease inhibitor cocktail) and centrifuged at 20,000g for 15 min. The supernatant (lysate) was used for Flag immunoprecipitation (IP) with M2 beads at 4 °C for 6 h. The beads were then washed in Flag IP wash buffer B (10 mM Tris-HCl (pH 7.5), 10 mM KCl, 1.5 mM MgCl₂, 500 mM NaCl, 0.1% Tween 20) five times before elution with the same buffer

containing Flag peptide. The eluate was filtered and concentrated using a 10-kDa cut-off concentrator, before aliquots were used for Coomassie blue staining and PIP Strip binding assay.

The binding assay was performed with PIP Strip (P-6001) (Echelon Biosciences) according to the protocol from the manufacturer. In brief, PIP Strip membrane was first blocked with PIP Strip Block Buffer (PBS with 0.1% Tween 20 and 3% fat-free BSA) for 1 h, before incubation with 2.5 μ g of purified target protein in fresh PIP Strip Block Buffer for another hour. The membrane was then washed with PIP Strip Wash Buffer (PBS with 0.1% Tween 20) three times (10 min each) and incubated with PIP Strip Wash Buffer containing Flag M2 antibody for 1 h. After that the membrane was washed three times before incubation with HRP mouse antibody in PIP Strip Wash Buffer for another hour. All the steps above were performed at room temperature. The PIP Strip membranes were then stained similar to regular immunoblotting. All samples were examined at the same time with similar exposure time. The experiment was repeated a total of three times with similar results.

Inducible recruitment of phospholipid phosphatases. Plasmids pTGN38-FRB-CFP and pmRFP-FKBP12-Sac1 (human, amino acids 2–516, with transmembrane domain removed) were provided by T. Balla (National Institutes of Health), whereas pmCherry-FKBP12-MTM1 (human) was ordered from Addgene (deposited by the laboratory of T. Balla). We also replaced Sac1 gene in pmRFP-FKBP12-Sac1 with other phosphatase genes to make pmRFP-FKBP12-Sac2 (human), pmRFP-FKBP12-lipin1 (human) and pmRFP-FKBP12-Fig4 (human).

In brief, we set up a COS-7 cell line stably expressing Flag-NLRP3. A single colony was selected to ensure homogenous expression of Flag-NLRP3 in all cells, and its behaviours (including stimulus-triggered dTGN formation, NLRP3 recruitment and activation) were confirmed to be similar to the original pooled cells. This stable cell line was then transiently transfected with both pTGN38-FRB-CFP and pmRFP(or mCherry)-FKBP12-phosphatase vectors (1 μ g/well of a 6-well plate for each vector) with Lipofectamine 2000 (Thermo Fisher Scientific) in OPTI-MEM medium. The medium was changed to regular medium 8 h after transfection and the cells were allowed to recover for another 2 h, before being split onto slides. After overnight culture, the cells were treated with rapamycin (1 μ M) in the absence or presence of nigericin (10 μ M) for 80 min before immunostaining with Flag antibody and Alexa Fluor 633 (pseudocoloured to green in images).

Statistics and reproducibility. Representative results from at least three experiments are shown for every figure except where specified otherwise in the figure legends.

For quantification of cells with the phenotype of interest, 25 non-overlapping whole-field images were randomly taken throughout the slide of each sample. Only the DAPI channel was used during the random selection of whole-field images to avoid bias in selection of cells with particular phenotypes. The number of cells with the phenotype of interest was recorded from 100 cells and this process was repeated three times. For quantification of TGN disassembly during the early time points (the first 30 min) of stimulation in primary wild-type BMDMs, 20 non-overlapping whole-field images were randomly taken throughout the slide for each sample and the number of TGN structures that were not connected with each other was quantified for 100 cells and grouped. Cells undergoing mitosis were excluded from the quantifications above, based on chromosomes morphology in DAPI channel, because these cells had mitotic Golgi disassembly. The sample sizes were selected based on power analysis of results from preliminary experiments, which shows that the sample sizes are not only sufficient to obtain desirable significance level (<0.01) and power (>90%), but also able to generate highly reproducible results with biological replicates.

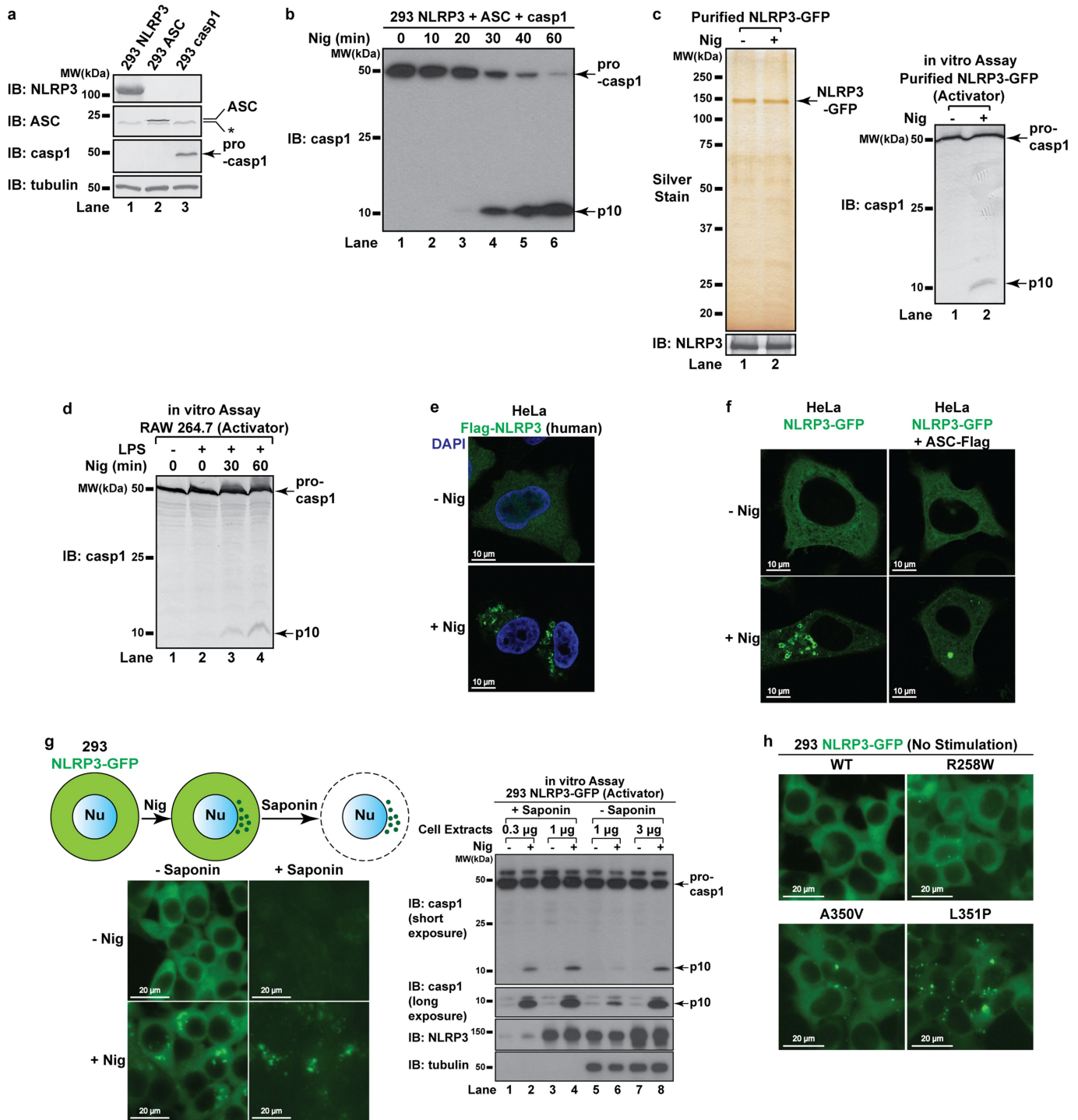
Data are presented as mean \pm s.d. Statistical analysis was performed using two-sided *t*-tests in GraphPad Prism 7. Statistical significance was determined with the Holm-Sidak method, with $\alpha = 0.01$. For co-localization analysis, images were selected similar to methods above, before Pearson's correlation coefficient (threshold regression, Costes) was calculated using the Coloc 2 plugin in ImageJ (v.1.51n)²⁸. For blinding, immunoblots and PIP Strip arrays were exposed with the same settings without the investigator knowing the order of the samples. They were identified later with markers assigned to them. Images were collected with randomization described above and a code number was assigned for each group. The investigator performed the quantification before the identities of the groups were added to the samples according to the code numbers.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

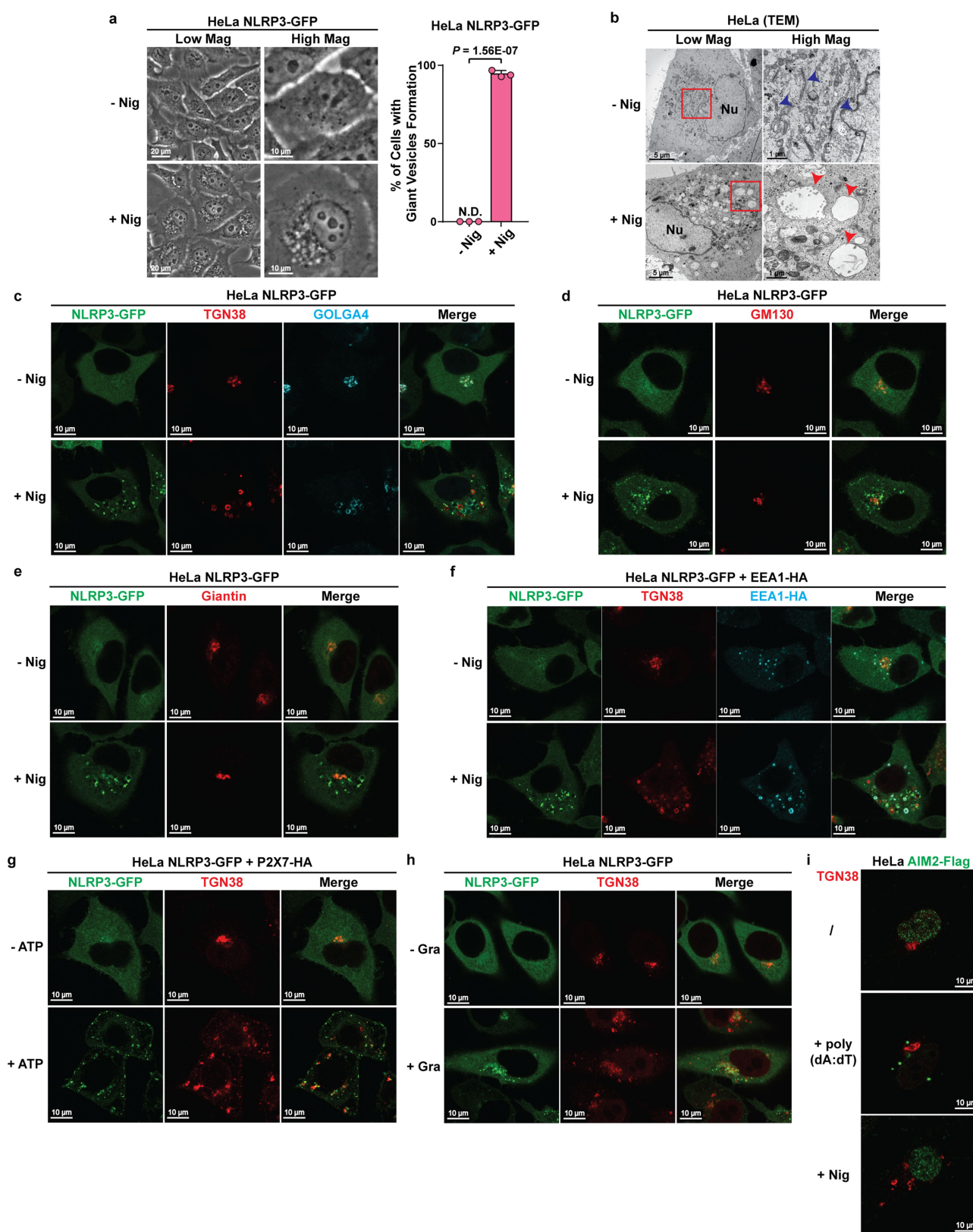
All important data generated or analysed during this study are included in this article. Additional supplementary data are available from the corresponding author upon request.

25. Bauernfeind, F. G. et al. Cutting edge: NF- κ B activating pattern recognition and cytokine receptors license NLRP3 inflammasome activation by regulating NLRP3 expression. *J. Immunol.* **183**, 787–791 (2009).
26. Tanaka, Y. & Chen, Z. J. STING specifies IRF3 phosphorylation by TBK1 in the cytosolic DNA signaling pathway. *Sci. Signal.* **5**, ra20 (2012).
27. Fernandes-Alnemri, T., Yu, J. W., Datta, P., Wu, J. & Alnemri, E. S. AIM2 activates the inflammasome and cell death in response to cytoplasmic DNA. *Nature* **458**, 509–513 (2009).
28. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).



Extended Data Fig. 1 | NLRP3 forms multiple puncta to activate the inflammasome pathway. a, Endogenous NLRP3, ASC and caspase-1 are not detectable in HEK-293T cells. Extracts from HEK-293T cell lines stably expressing the indicated proteins were examined by immunoblotting. *, non-specific band. **b**, Reconstitution of NLRP3 inflammasome pathway in HEK-293T cells. Cells stably expressing mouse NLRP3, ASC and caspase-1 were treated with nigericin (Nig) (10 μ M) for the indicated time, followed by immunoblotting. **c**, Highly purified NLRP3 showed signal-dependent activity in the in vitro assay. NLRP3-GFP was purified by fractionation and immunoprecipitation as detailed in Methods. Purity and activity were examined by silver staining (left) and the in vitro assay (right), respectively. **d**, The in vitro assay detects activation of endogenous NLRP3. RAW 264.7 cells were treated with LPS (50 ng ml⁻¹) for 3 h and stimulated with nigericin (10 μ M) for 60 min before cell extracts were collected for the in vitro NLRP3 activity assay.

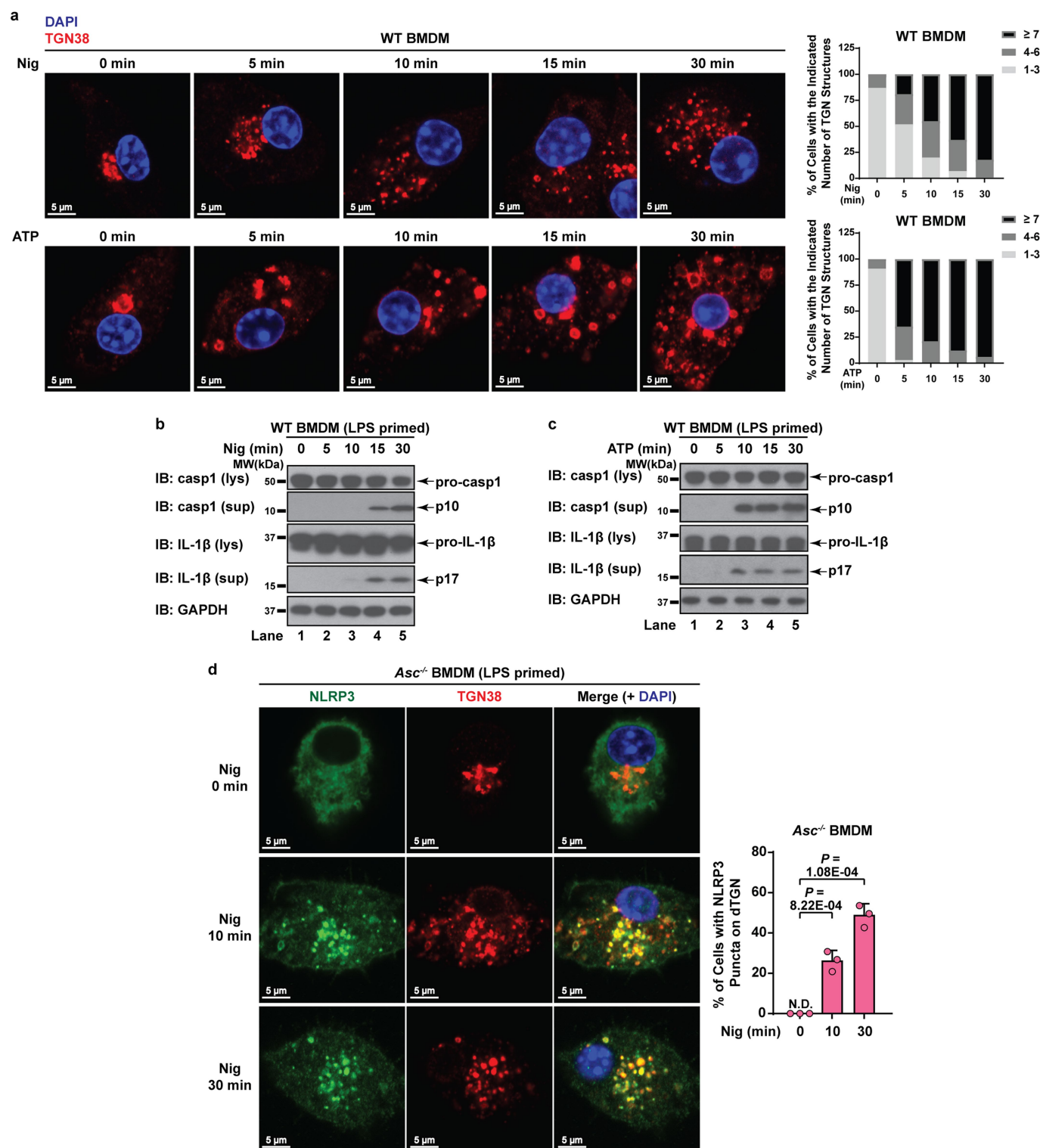
e, Human NLRP3 also formed puncta in response to nigericin. HeLa cells stably expressing Flag-NLRP3 were treated with nigericin (10 μ M) for 80 min before immunostaining with a Flag antibody. **f**, NLRP3 formed multiple puncta in the absence of ASC, and formed a large speck in the presence of ASC. Cells stably expressing the indicated proteins were treated as in **e** before imaging. **g**, NLRP3 puncta possessed high activity. Left, NLRP3 puncta induced by nigericin (10 μ M, 60 min) remained in the cells after saponin treatment. Nu, nucleus. Right, cell extracts with or without saponin treatment were examined by the in vitro assay. The p10 level in lane 4 is approximately 6.5 fold that in lane 6 based on quantification in imageJ (normalization by NLRP3 band intensity). Only activator cell extracts were used for tubulin immunoblot. **h**, Constitutively active mutants of NLRP3 formed puncta without stimulation. Cells stably expressing the indicated proteins were imaged in the absence of stimulation.



Extended Data Fig. 2 | See next page for caption.

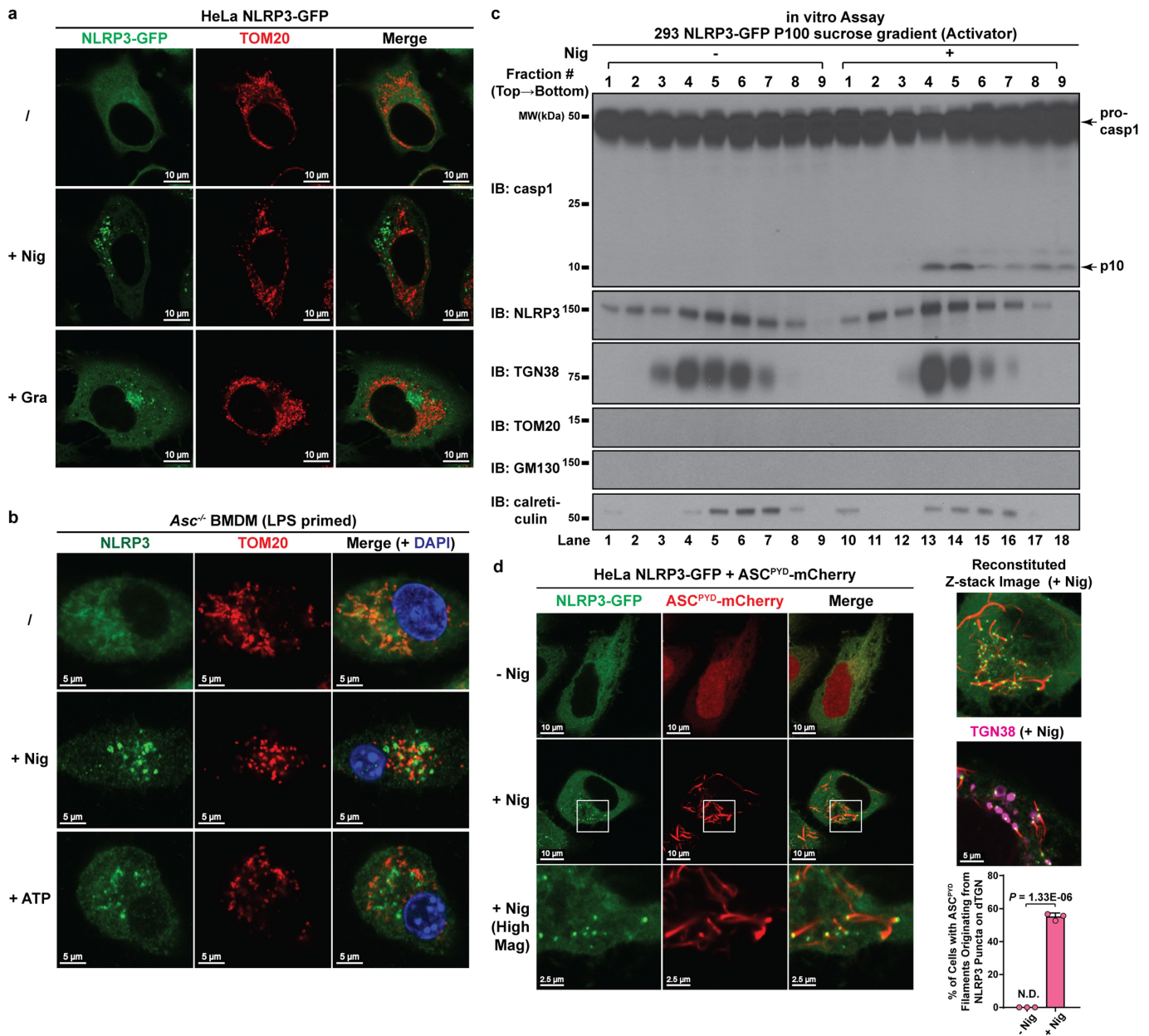
Extended Data Fig. 2 | NLRP3 aggregates on stimulus-triggered dTGN. **a**, Nigericin treatment induced formation of giant vesicles in the perinuclear region. HeLa cells expressing NLRP3–GFP treated with nigericin (10 μ M) for 80 min were examined with phase-contrast microscopy. Mag, magnification. Cells with giant vesicle formation was quantified from 100 cells ($n = 3$, mean \pm s.d., two-sided t -test). N.D., not detectable. **b**, Ultrastructural analysis of nigericin-induced dispersed TGN (dTGN) vesicles. HeLa cells treated as in **a** were examined by transmission electron microscopy. Blue arrowheads indicate Golgi stacks under resting conditions and red arrowheads indicate nigericin-induced dTGN vesicles. Representative images from two independent experiments (more than 30 cells were examined for each condition in each experiment) are shown. **c**, Nigericin triggered the formation of dTGN, on which NLRP3 aggregated. HeLa cells stably expressing the indicated protein were stimulated as in **a** before immunostaining for the TGN markers TGN38 and GOLGA4. **d**, **e**, *cis* and medial Golgi remained intact after nigericin

treatment. Cells treated as in **a** were immunostained for GM130 (*cis*-Golgi marker) or giantin (*cis*- and medial-Golgi marker). **f**, NLRP3 aggregated on dispersed TGN38-positive vesicles, some of which were also EEA1-positive. HeLa cells stably expressing NLRP3–GFP and EEA1–HA were treated as in **a** before immunostaining for TGN38 and HA. **g**, ATP stimulation led to NLRP3 aggregation on dTGN. HeLa cells stably expressing NLRP3–GFP and P2X₇–HA were treated with ATP (5 mM) for 80 min before imaging. P2X₇ is a purinergic receptor that is essential for ATP-mediated NLRP3 inflammasome activation. **h**, Stimulation with gramicidin led to NLRP3 aggregation on dTGN. HeLa cells were treated with gramicidin (5 μ M) for 80 min before imaging. **i**, DNA stimulation does not cause TGN dispersion or AIM2 recruitment to TGN. HeLa cells stably expressing AIM2–Flag were mock-transfected, transfected with poly(dA:dT) (1.5 μ g ml^{−1}) for 3 h or incubated with nigericin (10 μ M) for 80 min before immunostaining with antibodies against Flag (AIM2–Flag) or TGN38.



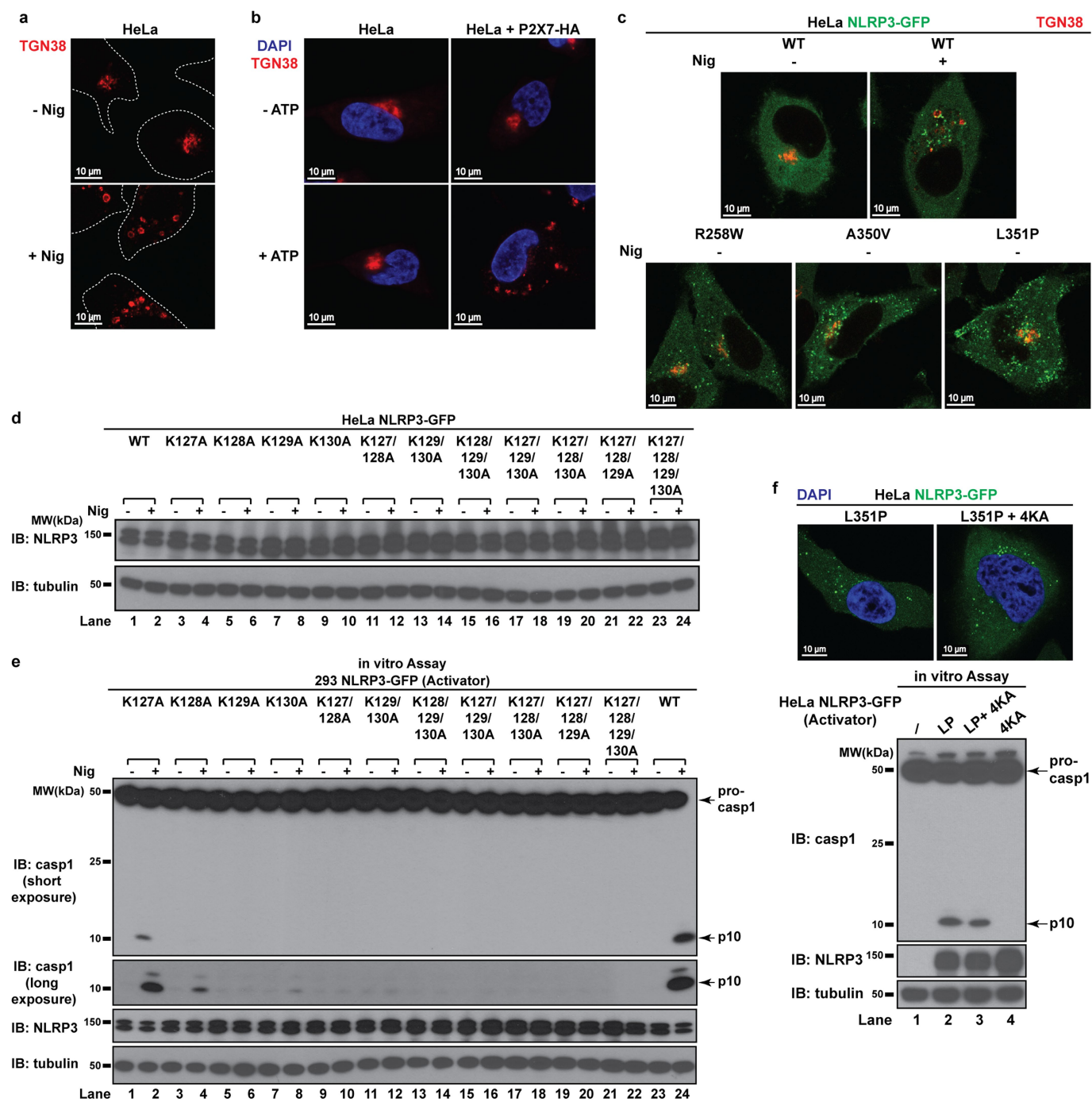
Extended Data Fig. 3 | Endogenous NLRP3 is recruited to dTGN in primary macrophages. **a**, Substantial TGN disassembly occurred at early time points in wild-type BMDMs. Cells were primed with LPS (50 ng ml^{-1}) for 3 h, followed by stimulation with nigericin ($10 \text{ }\mu\text{M}$) or ATP (5 mM) for the indicated time, and immunostained for TGN38. Right, to quantify TGN disassembly, the numbers of TGN structures not connected with each other for each cell were quantified from 100 randomly selected cells and grouped as shown. **b**, **c**, Nigericin-induced cleavage of caspase-1 and IL-1 β did not occur until 15 min (for nigericin)

or 10 min (for ATP) after stimulation in wild-type BMDMs. Cells were treated as in **a** before lysates were collected for immunoblotting. **d**, Endogenous NLRP3 aggregation on dTGN could be detected as early as 10 min after nigericin treatment in ASC-deficient BMDMs. Cells were primed with LPS (50 ng ml^{-1}) for 3 h, followed by nigericin ($10 \text{ }\mu\text{M}$) treatment for 0, 10 or 30 min before imaging. Cells with NLRP3 puncta on dTGN were quantified from 100 cells ($n = 3$, mean \pm s.d., two-sided t -test).



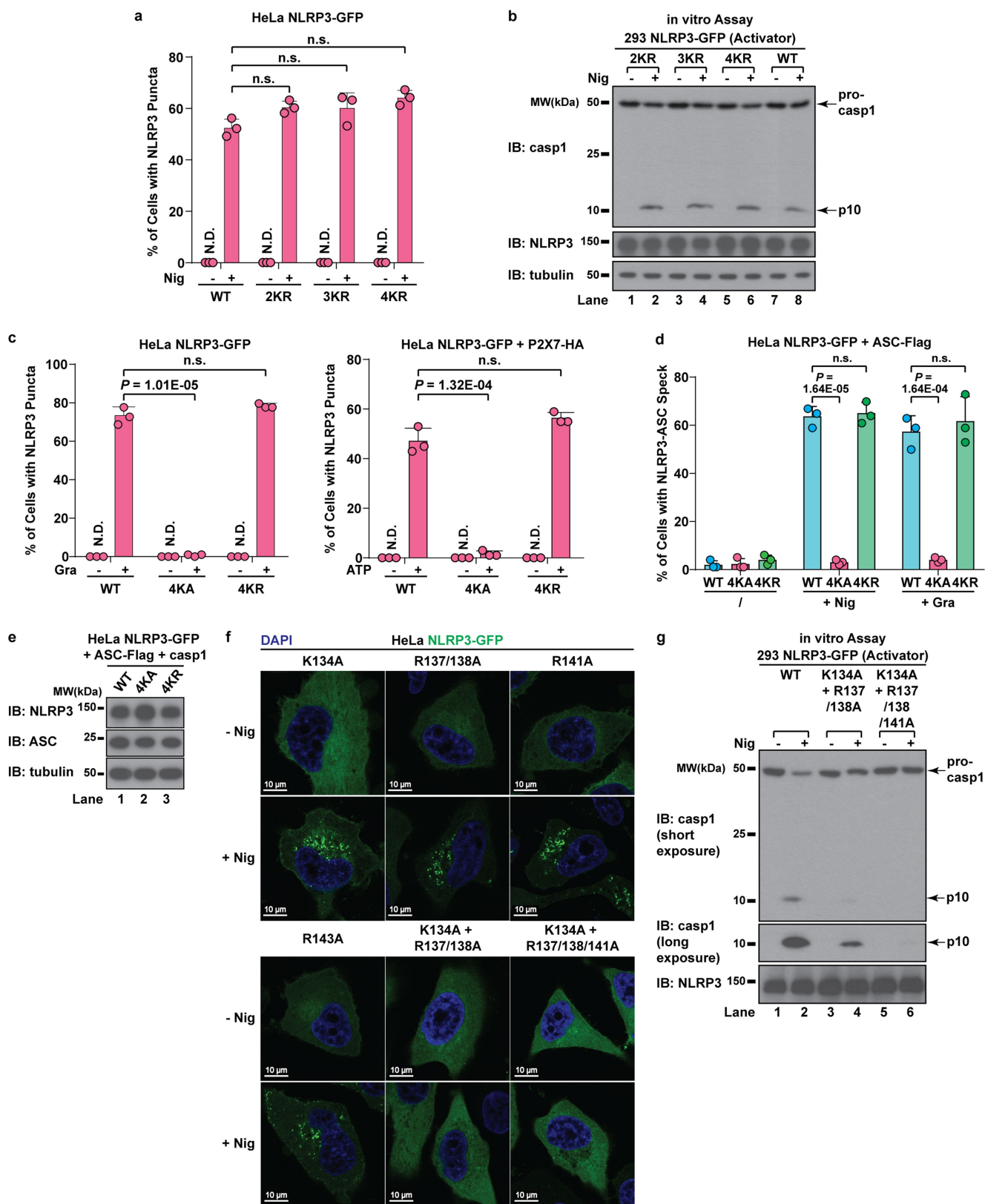
Extended Data Fig. 4 | NLRP3 activity is strongly associated with dTGN but not mitochondria. **a**, NLRP3 did not translocate to mitochondria upon stimulation in HeLa cells. HeLa cells expressing NLRP3-GFP were stimulated with nigericin (10 μ M) or gramicidin (5 μ M) for 80 min and were immunostained for TOM20 (mitochondrial marker). **b**, Neither nigericin- nor ATP-induced NLRP3 puncta were co-localized with mitochondria in ASC-deficient BMDMs. Cells were primed with LPS (50 ng ml⁻¹) for 3 h, followed by nigericin (10 μ M) or ATP (5 mM) treatment for 60 min before immunostaining for endogenous NLRP3 and TOM20. **c**, NLRP3 activity in P100 (light membrane) fraction was strongly associated with dTGN but not mitochondria. P100 fraction collected from Fig. 1c was separated by sucrose gradient

ultracentrifugation. Fractions were collected and tested for activity in the in vitro NLRP3 activity assay (top). TOM20 (mitochondrial marker) and GM130 (*cis*-Golgi marker) were not detectable on immunoblots even after prolonged exposure. **d**, dTGN-localized NLRP3 puncta can initiate aggregation of ASC-PYD. HeLa cells stably expressing the indicated proteins were incubated with nigericin (10 μ M) for 80 min before imaging. ASC-PYD, residues 1–90 of mouse ASC. Right, top to bottom: reconstituted z-stack image of a representative nigericin-treated cell; nigericin-treated cell co-immunostained for TGN38 (pseudocoloured in magenta); cells with ASC-PYD filaments originating from dTGN-localized NLRP3 puncta were quantified from 100 cells ($n = 3$, mean \pm s.d., two-sided *t*-test).



Extended Data Fig. 5 | The polybasic region mediates dTGN recruitment and activation of NLRP3. **a**, NLRP3 is not required for dTGN formation in response to nigericin in HeLa cells. HeLa cells (without NLRP3 expression) were treated with nigericin (10 μ M) for 80 min and immunostained for TGN38. Cell borders are outlined with dashed lines. **b**, ATP induced dTGN formation in a manner dependent on P2X₇ but not NLRP3. HeLa cells or HeLa cells stably expressing P2X₇-HA were incubated with ATP (5 mM) for 80 min before immunostaining for TGN38. **c**, Constitutively active mutants of NLRP3 can bypass the TGN-recruitment step by spontaneously forming aggregates in the cytosol.

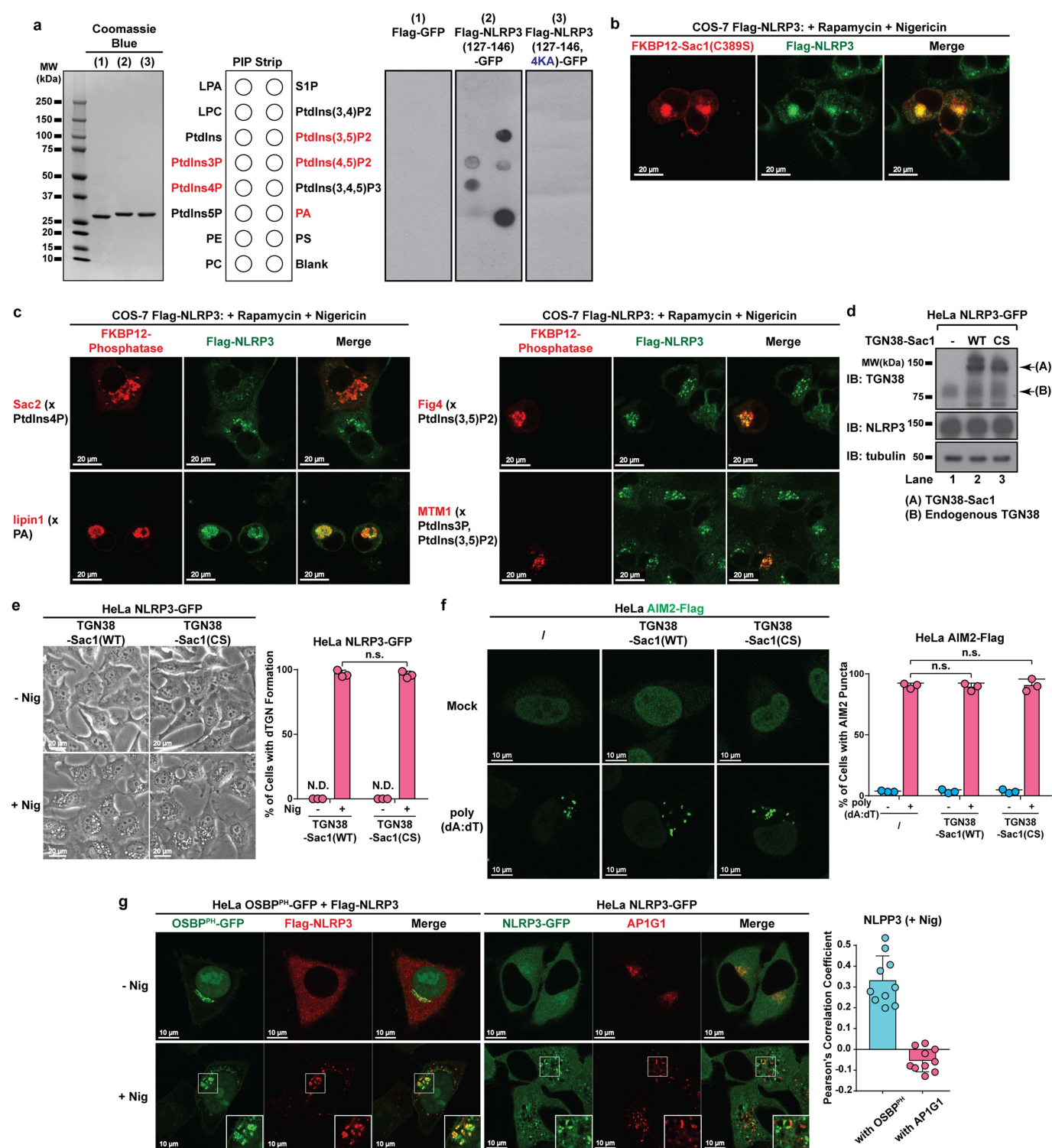
Cells stably expressing the indicated proteins were untreated or treated with nigericin (10 μ M) for 80 min before immunostaining for TGN38. **d**, Immunoblots of cells used in Fig. 3c. **e**, The KKKK motif is critical for nigericin-induced NLRP3 activation. Cells stably expressing the indicated proteins were untreated or treated with nigericin (10 μ M) for 60 min before being tested in the *in vitro* NLRP3 activation assay. **f**, Mutations in the KKKK motif do not compromise the ability of constitutively active NLRP3 to activate caspase-1. Cells stably expressing the indicated proteins were examined by fluorescence microscopy (top) and the *in vitro* NLRP3 activity assay (bottom) without stimulation. LP, NLRP3(L351P).



Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | The polybasic region of NLRP3 functions through its positive charge. **a**, Mutations of the KKKK motif to arginine do not affect nigericin-induced NLRP3 puncta formation. HeLa cells stably expressing the indicated proteins were treated with nigericin (10 μ M) for 80 min before imaging. Cells with NLRP3 puncta were quantified from 100 cells ($n = 3$, mean \pm s.d., two-sided t -test). n.s., not significant ($\alpha = 0.01$). 2KR, NLRP3(K127R/K128R); 3KR, NLRP3(K127R/K128R/K129R); 4KR(K127R/K128R/K129R/K130R). **b**, Mutations of the KKKK motif to arginine do not impair nigericin-induced NLRP3 activation. Cells stably expressing the indicated proteins were treated with nigericin (10 μ M) for 60 min before cell extracts were examined in the in vitro NLRP3 activity assay. **c**, The positive charge of the KKKK motif is important for gramicidin- and ATP-induced NLRP3 puncta

formation. HeLa cells stably expressing the indicated proteins were treated with gramicidin (5 μ M) (left) or ATP (5 mM) (right) for 80 min, before the percentage of cells with NLRP3 puncta was analysed as in **a**. **d**, The positive charge of the KKKK motif is essential for NLRP3 to polymerize ASC. Cells stably expressing the indicated proteins were treated with nigericin (10 μ M) or gramicidin (5 μ M) for 80 min before immunostaining for both NLRP3 and ASC. The percentage of cells with NLRP3–ASC specks was analysed as in **a**. **e**, Immunoblots of cells used in Fig. 3d. **f**, **g**, The second positively charged region of NLRP3 is also important for its aggregation on dTGN and activation in a manner dependent on the number of remaining positively charged residues. Cells stably expressing the indicated proteins were treated with nigericin (10 μ M) before imaging (**f**) or examined by the in vitro NLRP3 activity assay (**g**).

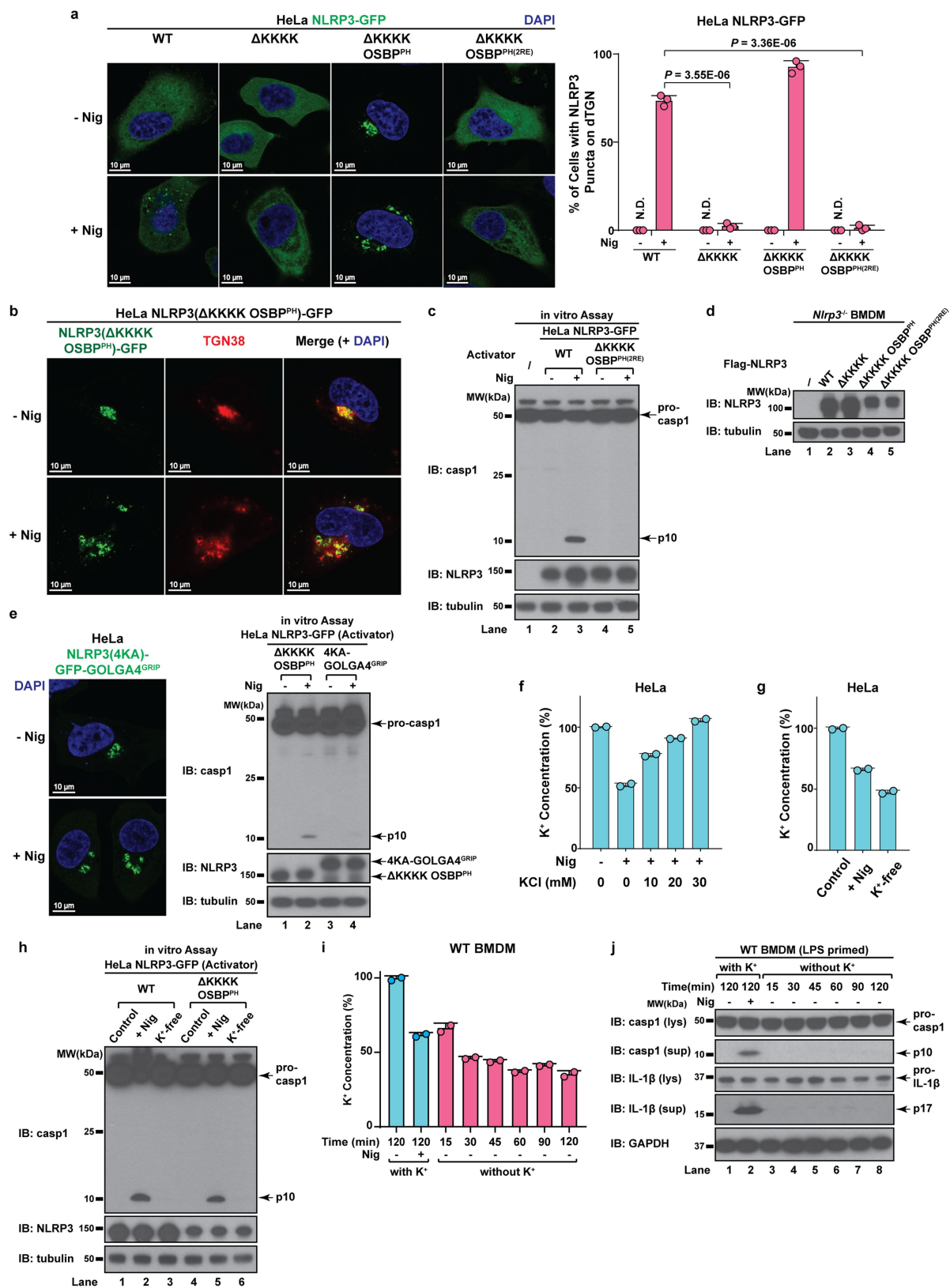


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | NLRP3 is recruited to dTGN via binding to

PtdIns4P. **a**, The polybasic region of NLRP3 interacts with phospholipids in vitro through its positive charge. Left, Coomassie blue staining of purified proteins of interest, with Flag-GFP as a control. Right, PIP Strip membranes blotted with various lipids were incubated sequentially with proteins of interest, Flag antibody and HRP secondary antibody before exposure. Phospholipids with positive binding on the second PIP Strip are highlighted in red. **b**, Catalytically inactive Sac1 did not impair the recruitment of NLRP3 to dTGN. COS-7 cells stably expressing Flag-NLRP3 were transiently transfected with TGN38-FRB and mRFP-FKBP12-Sac1(C389S), incubated with rapamycin (1 μ M) and nigericin (10 μ M) for 80 min before imaging. Note that nigericin-induced dTGN in COS-7 cells sometimes looks like a single cluster because of the relatively small size of COS-7 cells and the high intensity of fluorescence signal. Under the phase contrast microscope these dTGN particles were observed to be individual vesicles adjacent to each other (data not shown). **c**, Only PtdIns4P phosphatase blocked NLRP3 recruitment to dTGN. Similar to

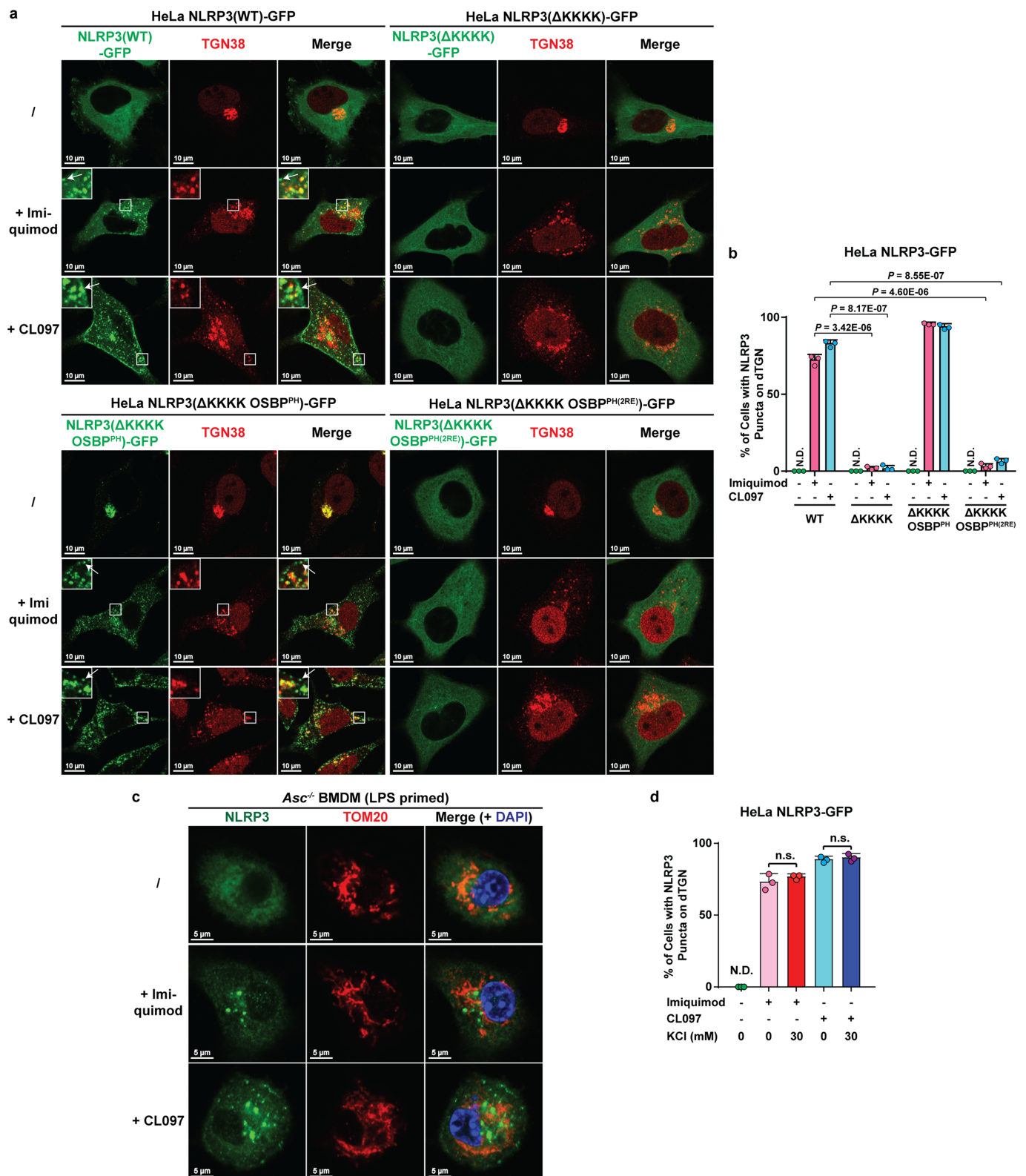
b, except that indicated phosphatases were used. The target phospholipids are labelled after 'x'. **d**, Immunoblots of HeLa NLRP3-GFP cells stably expressing TGN38-Sac1 or TGN38-Sac1(C389S). **e**, TGN-targeted Sac1 did not affect general cell morphology or nigericin-induced dTGN formation. Cells were treated with nigericin (10 μ M) for 80 min before imaging with phase contrast microscopy. Cells with dTGN formation were quantified from 100 cells ($n = 3$, mean \pm s.d., two-sided t -test). n.s., not significant ($\alpha = 0.01$). **f**, TGN-targeted Sac1 did not affect AIM2 aggregation. Cells stably expressing the indicated proteins were transfected with poly(dA:dT) (1.5 μ g ml⁻¹) for 3 h before imaging. The percentage of cells with AIM2 aggregates was analysed as in **e**. **g**, NLRP3 puncta were restricted to PtdIns4P-enriched microdomains. Cells stably expressing the indicated proteins were treated with nigericin (10 μ M) for 80 min before imaging. Inset, higher magnification of the dTGN. Co-localization analysis was performed by calculating Pearson's correlation coefficient (threshold regression, Costes) using the Coloc 2 plugin of ImageJ. Data are presented as mean \pm s.d.



Extended Data Fig. 8 | See next page for caption.

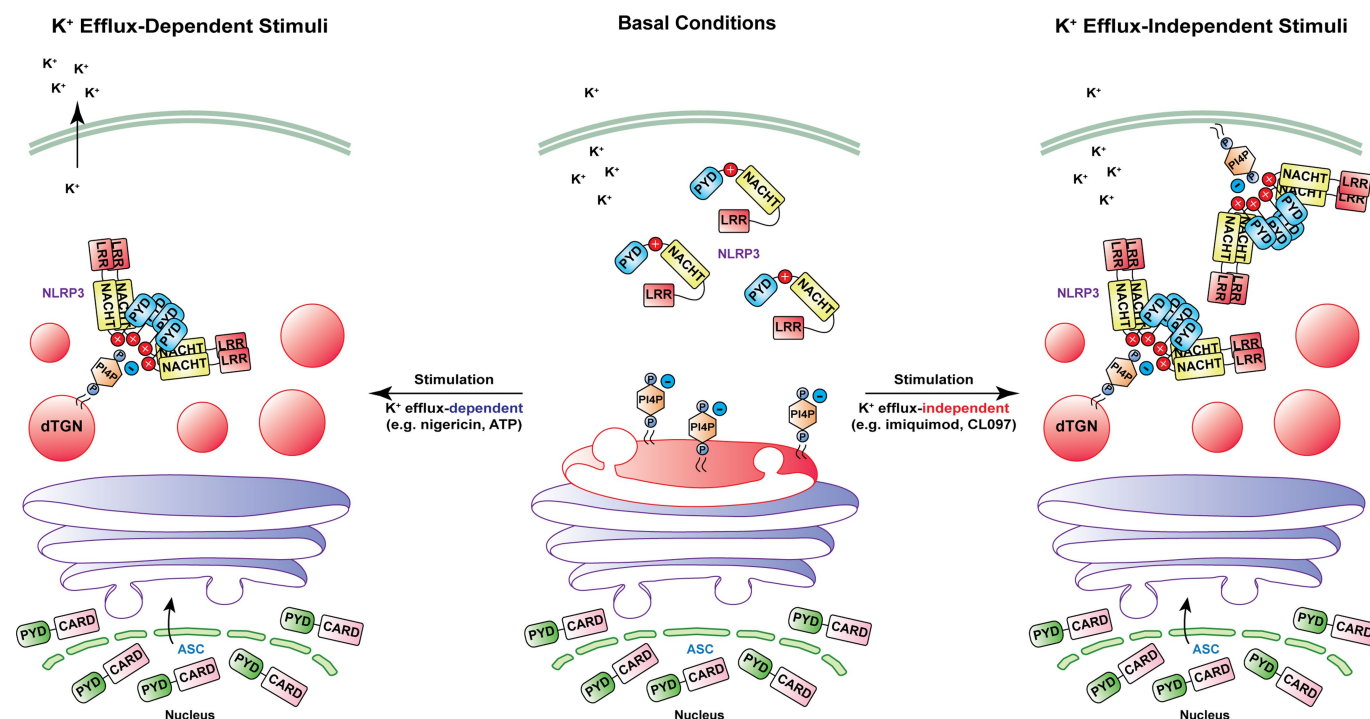
Extended Data Fig. 8 | Binding to PtdIns4P on dTGN is essential for NLRP3 inflammasome activation. **a, b,** The KKKK motif of NLRP3 can be functionally replaced with a PtdIns4P-binding domain of OSBP (OSBP-PH). **a,** Cells stably expressing the indicated proteins were treated with nigericin (10 μ M) for 80 min before imaging. Cells with NLRP3 on dTGN were quantified from 100 cells ($n = 3$, mean \pm s.d., two-sided t -test). 2RE, NLRP3(R107/108E). Note that NLRP3(Δ KKKK OSBP-PH) that constitutively localized on intact TGN without stimulation was not counted in the quantification. **b,** Replacement of the KKKK motif with OSBP-PH allowed NLRP3 to be constitutively localized on TGN. Cells were treated as in **a** before immunostaining for TGN38. **c,** Mutations of OSBP-PH domain that abrogate its binding to PtdIns4P also abolish its ability to functionally rescue NLRP3(Δ KKKK). Cells were treated as in **a** and cell extracts were examined by the in vitro NLRP3 activity assay. **d,** Immunoblots for primary NLRP3-deficient BMDMs rescued with Flag-NLRP3 (wild type or mutants), which were used for experiment in Figs. 5b, 6d. The cells were infected with lentivirus encoding the indicated proteins for 6 days before immunoblotting. **e,** Recruitment of NLRP3 to non-PtdIns4P-enriched regions of TGN is not sufficient to support its activation. Cells stably expressing the indicated proteins were treated as in **a** before examination by fluorescence microscopy (left; images for NLRP3(Δ KKKK OSBP-PH) are shown in **a**) and the in vitro NLRP3 assay

(right). **f,** Extracellular KCl at 30 mM was sufficient to completely block nigericin-induced K^+ efflux. Cells were treated with nigericin (10 μ M) for 80 min in the presence of increasing concentrations of KCl, and cell extracts were collected for measurement of intracellular K^+ concentration (shown as percentage change compared to untreated sample, mean \pm s.d.). Representative results from two independent experiments (each containing two samples for each condition) are shown. **g,** Incubation in K^+ -free medium induced spontaneous K^+ efflux. Cells were incubated in Hanks' buffer containing 5 mM K^+ for the first two conditions, or Hanks' buffer in which K^+ was replaced by Na^+ for the third condition (K^+ -free). The second condition also contained nigericin (10 μ M). After 80 min, the cell extracts were collected for intracellular K^+ measurement using methods similar to **f**. **h,** K^+ efflux alone is not sufficient to activate either wild-type NLRP3 or NLRP3(Δ KKKK OSBP-PH). Cells stably expressing the indicated proteins were treated as in **g** before testing with the in vitro NLRP3 activity assay. **i,** Incubation in K^+ -free medium induced spontaneous K^+ efflux in primary wild-type BMDMs. Cells were primed with LPS (50 ng ml $^{-1}$) for 3 h, before treatment as in **g** for the indicated time period. Intracellular K^+ concentrations were then measured and analysed as in **f**. **j,** K^+ efflux alone is not sufficient to activate endogenous NLRP3 in primary wild-type BMDMs. Cells treated as in **i** were used for immunoblotting. Lys, lysate; sup, supernatant.



Extended Data Fig. 9 | K^+ efflux-independent stimuli also induced TGN dispersion and PtdIns4P-dependent NLRP3 recruitment. **a, b,** K^+ efflux-independent stimuli also induced NLRP3 aggregation on dTGN through the KKKK motif. HeLa cells stably expressing the indicated proteins were treated with imiquimod or CL097 (45 μ g ml⁻¹) for 80 min before imaging. High magnification images are shown in the inset. Arrows indicate representative plasma-membrane-localized NLRP3 puncta, which were separated from TGN38-positive compartments owing to the partial separation of PtdIns4P and TGN38. Cells with NLRP3 puncta on dTGN were quantified from 100 cells ($n = 3$, mean \pm s.d.; two-sided t -test).

c, Neither imiquimod- nor CL097-induced NLRP3 puncta were co-localized with mitochondria in ASC-deficient BMDMs. Cells were primed with LPS (50 ng ml⁻¹) for 3 h and incubated with imiquimod or CL097 (45 μ g ml⁻¹) for 60 min, before immunostaining for endogenous NLRP3 and TOM20 (mitochondrial marker). **d,** High extracellular KCl had no effect on imiquimod- or CL097-induced NLRP3 aggregation on dTGN. HeLa cells expressing NLRP3-GFP cells were treated with imiquimod or CL097 (45 μ g ml⁻¹) in the presence of KCl (0 or 30 mM) for 80 min before imaging. Results were analysed as in **b**. n.s., not significant ($\alpha = 0.01$).



Extended Data Fig. 10 | Model: NLRP3 aggregation on dTGN through PtdIns4P binding is a common cellular signal essential for the inflammasome activation by diverse stimuli. Centre, under basal conditions, NLRP3 is diffused in the cytosol and TGN (red) remains as a single cluster attached to medial- and *cis*-Golgi (purple). Left, when cells are stimulated with K⁺ efflux-dependent stimuli such as nigericin or ATP, TGN is disassembled into multiple dispersed structures (dTGN), whereas *cis*- and medial-Golgi stacks remain intact. These stimuli also trigger K⁺ efflux, which helps recruit NLRP3 to dTGN via ionic bonding between negatively charged PtdIns4P on dTGN membranes and the positively charged polybasic region of NLRP3. Right, when cells are stimulated

with K⁺ efflux-independent stimuli such as imiquimod and CL097, TGN is also disassembled, although more drastically. PtdIns4P is partially separated from other TGN compartments; some may be enriched on the plasma membrane. NLRP3 is then recruited to these PtdIns4P-containing microdomains through its polybasic region. For both types of stimulation, dTGN serves as a scaffold for NLRP3 to aggregate in the form of multiple puncta, which then interact with ASC to activate the downstream signalling cascade. This model shows that aggregation of NLRP3 on dTGN through PtdIns4P binding is a common cellular signal that is essential for its activation by diverse stimuli.

Type 9 secretion system structures reveal a new protein transport mechanism

Frédéric Lauber^{1,4}, Justin C. Deme^{2,3,4}, Susan M. Lea^{2,3*} & Ben C. Berks^{1*}

The type 9 secretion system (T9SS) is the protein export pathway of bacteria of the Gram-negative Fibrobacteres–Chlorobi–Bacteroidetes superphylum and is an essential determinant of pathogenicity in severe periodontal disease. The central element of the T9SS is a so-far uncharacterized protein-conducting translocon located in the bacterial outer membrane. Here, using cryo-electron microscopy, we provide structural evidence that the translocon is the T9SS protein SprA. SprA forms an extremely large (36-strand) single polypeptide transmembrane β -barrel. The barrel pore is capped on the extracellular end, but has a lateral opening to the external membrane surface. Structures of SprA bound to different components of the T9SS show that partner proteins control access to the lateral opening and to the periplasmic end of the pore. Our results identify a protein transporter with a distinctive architecture that uses an alternating access mechanism in which the two ends of the protein-conducting channel are open at different times.

The T9SS is responsible for protein export across the outer membrane of bacteria of the Fibrobacteres–Chlorobi–Bacteroidetes (FCB) superphylum^{1,2}. In *Porphyromonas gingivalis* and related human oral pathogens, which cause severe periodontal disease, the T9SS is an essential determinant of virulence². The T9SS is also a pathogenicity factor in major bacterial diseases of farmed fish, including columnaris disease and cold-water disease^{3,4}. Bacteroidetes exhibit a unique and rapid gliding motility in which cell surface adhesins move on helical tracks⁵. This gliding motility is intimately linked to the T9SS through the involvement of the T9SS in adhesin export and possibly through the use of a shared motor complex⁶.

T9SS substrates are large (100–650 kDa) multi-domain proteins that fold in the periplasm before being exported through an outer membrane translocon. Substrates are targeted to the translocon by means of an approximately 100-amino-acid C-terminal folded signal domain (CTD)^{7,8}. Distinct type A and type B CTDs have been defined and these differ in the T9SS components required for their export^{1,8,9}.

The Bacteroidete T9SS contains at least 15 proteins^{1,2,10}. GldK, GldL, GldM and GldN form a trans-periplasmic motor complex that is thought to use the inner membrane protonmotive force to drive secretion at the outer membrane^{11,12}. PorQ, PorU, PorV and PorZ form a cell surface-exposed attachment complex¹³ that proteolytically removes the substrate CTD following transport¹⁴ and covalently links some substrate proteins to a lipopolysaccharide anchor¹⁵. A pool of PorV proteins in the outer membrane is thought to shuttle substrate proteins from the translocon to the attachment complex¹³. The functions of the remaining T9SS components are unclear. In particular, the identities of the components that form the outer membrane translocon itself are unknown.

We reasoned that the protein-conducting channel of the T9SS translocon must be formed by a protein that traverses the outer membrane. However, most of the outer membrane-spanning T9SS components can be predicted to form β -barrels of 14 or fewer strands, with internal pores that would be too narrow to accommodate the folded T9SS substrate proteins. The only exception is the protein SprA (also termed Sov), a 267-kDa polypeptide with no sequence similarity to other currently identified proteins^{16,17}. This logic led us to hypothesize that SprA

forms the protein-conducting channel of the T9SS translocon. In support of this contention, we note that the protein that forms the transport channel would also be expected to be one of the five Bacteroidete T9SS components (SprA, SprE, PorU, PorV, and PorZ) that are retained in a minimal T9SS found in other members of the FCB superphylum (ref.¹⁸ and our analysis).

Characterization of SprA

We constructed strains of the gliding bacterium *Flavobacterium johnsoniae* in which the SprA protein was fused to either a HaloTag domain, to allow fluorophore labelling, or a Twin-Strep tag, to permit affinity purification. These strains retain full T9SS function (Extended Data Fig. 1a–d and Supplementary Video 1).

The fluorophore-labelled SprA fusion protein localized to an average of 7 ± 2 (s.d.) well-resolved foci per cell (Fig. 1a, b). Photobleaching traces showed that most foci contained a single SprA molecule (1.2 ± 0.4 (s.d.) photobleaching steps per focus, Fig. 1c). The SprA foci are stationary with respect to the cell body in both immobile and gliding cells (Supplementary Video 2). Thus, SprA is not physically linked to the mobile cell surface adhesins.

We used the Twin-Strep-tagged SprA fusion protein to isolate SprA from otherwise native cells. Proteomics analysis of the SprA preparation showed that the majority of the purified SprA molecules were proteolytically clipped between residues 1905 and 1946 (Fig. 1d and Extended Data Fig. 1e). We found that SprA co-purifies with the known T9SS component PorV as well as two additional proteins, Fjoh_1759 and Fjoh_4997 (Fig. 1d). Fjoh_4997 is a lipoprotein peptidyl-prolyl *cis-trans* isomerase of the FKBP superfamily, hereafter referred to as PPI. Fjoh_1759 is a protein of unknown function but with a phylogenetic distribution that strongly correlates with the presence of a T9SS¹⁰. For reasons set out below, we term Fjoh_1759 the Plug protein.

We used cryo-electron microscopy (cryo-EM) to determine the structures of two distinct complexes in which SprA associates either with the PPI and PorV proteins (PorV complex) or the PPI and Plug proteins (Plug complex; Fig. 1e–g, Extended Data Table 1, Extended Data Fig. 2, Supplementary Videos 3, 4). Fourier shell correlation (FSC) defined the resolution of the PorV complex as 3.5 Å overall (local

¹Department of Biochemistry, University of Oxford, Oxford, UK. ²Sir William Dunn School of Pathology, University of Oxford, Oxford, UK. ³The Central Oxford Structural Molecular Imaging Centre (COSMIC), University of Oxford, Oxford, UK. ⁴These authors contributed equally: Frédéric Lauber, Justin C. Deme. *e-mail: susan.lea@path.ox.ac.uk; ben.berks@bioch.ox.ac.uk

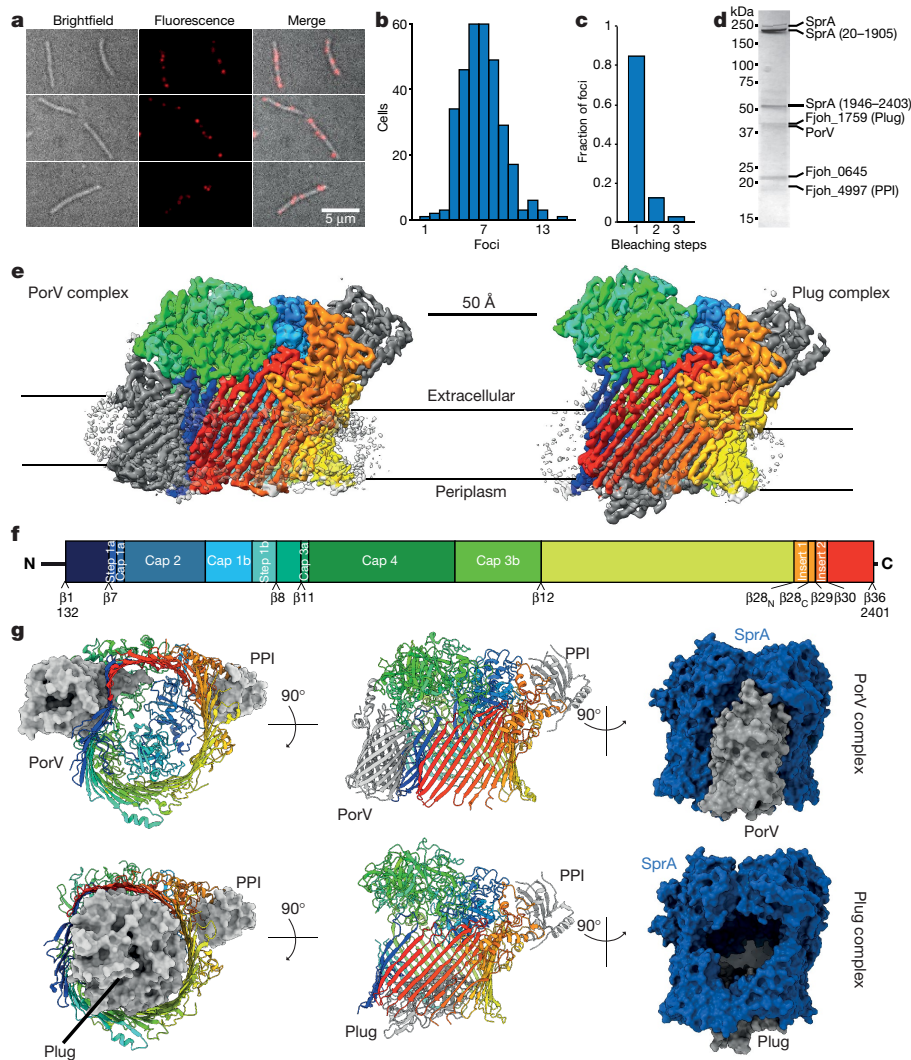


Fig. 1 | Characterization of SprA. **a**, Localization of fluorophore-labelled SprA in *F. johnsoniae* cells. Exemplar data from experiments repeated on three different cultures with similar results. **b**, **c**, Population distributions of SprA foci per cell (from 314 cells) (**b**) and SprA molecules per focus (**c**) from 119 photobleaching traces. **d**, Coomassie-stained SDS-PAGE gel of affinity-purified SprA with band identities from peptide mass spectrometry. Fjoh_0645 is a biotin-containing contaminant protein that

resolution varying from 3.3 to 5.0 Å) whereas the Plug complex is at a resolution of 3.7 Å (local resolution 3.4–4.7 Å; Extended Data Fig. 2c, d). For both complexes, an α -carbon trace was initially built through the electron microscopy (EM) volume. Regions corresponding to PorV, PPI, and Plug were assigned by dissecting the EM volume into domains and searching the Protein Data Bank (PDB; <https://www.rcsb.org/>) for structural homologues (Extended Data Fig. 3a). Sequences could be docked into these regions of the structures guided by the identified homologues. Sequence docking was then extended manually into SprA (Extended Data Fig. 2e). With the exception of about 100 amino acids at the N terminus of SprA and some short disordered loops, the chains were continuously traced.

Both SprA and PorV form transmembrane β -barrels (Fig. 1g). This allowed us to assign the orientation of the SprA complexes in the outer membrane using the principle that the last β -strand of bacterial single subunit β -barrel proteins is always at the periplasmic side of the membrane¹⁹. The site at which proteolytic clipping occurs during purification of SprA falls within a 50-residue loop that extends from the periplasmic side of the SprA barrel between strands 24 and 25. This loop is not visible in the EM density for either of our two complexes.

binds to the streptavidin affinity matrix. Similar data were obtained for three independent preparations. For gel source data, see Supplementary Fig. 1. **e**, EM density of the SprA complexes. Membrane position inferred from the location of the detergent micelles. **f**, Domain organization of SprA. **g**, Overall structures of the SprA complexes. The same SprA domain colouring scheme is applied in **e–g** except in **g**, far right panel (blue, SprA; grey, PorV or Plug).

SprA architecture

Using the location of the detergent associated with the complexes to define the likely plane of the membrane, we find that the SprA barrel extends about 20 Å above the membrane on the extracellular side of the complex (Fig. 1e). Two large, folded inserts between barrel strands 7–8 and 11–12 form a 50 Å-high cap structure that seals the extracellular end of the barrel. Each of the two cap inserts is composed of two domains of predominantly β -sheet structure, with the second domain inserted between the first and second halves of the first domain (Figs. 1f, g, 2a, b). In the PorV complex, the polypeptide sequences that connect the terminus of the first cap domain to the barrel fold together to create another split- β domain that forms a step-like structure running down the inside of the SprA barrel (Figs. 2a, b, 3a). This step domain shows high surface sequence conservation (Fig. 3c), indicating that it may participate in a functional interaction with a protein that is not present in our purified SprA complexes. The equivalent EM density is not well defined in the Plug complex. With the exception of the step structure, the inside of the SprA barrel is empty (Fig. 3c, Supplementary Video 3), resulting in an internal solvent-filled pore of approximately 70 Å in diameter (measured C α to C α ; Fig. 3a, Extended Data Fig. 3b). A pore of this size would be large enough to permit the passage of the folded substrates of

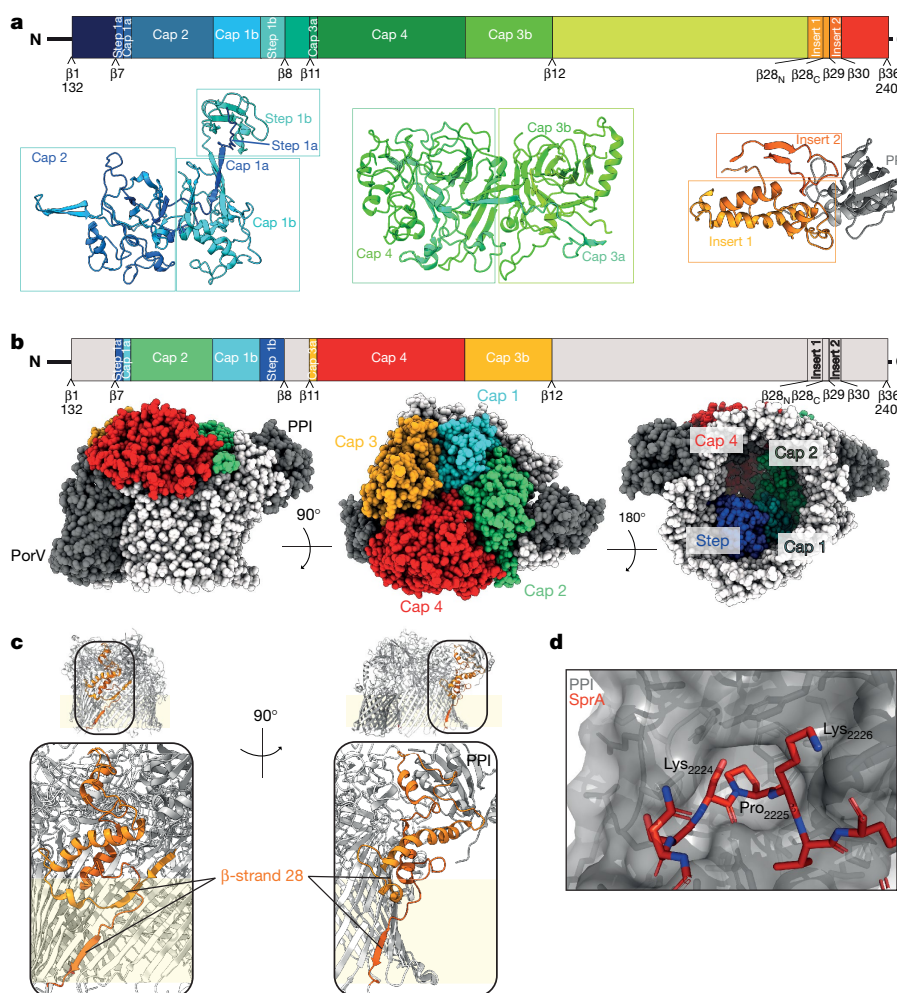


Fig. 2 | Structural features of SprA extracellular domains. a, Structures of the SprA step domain, cap domain, and insert sequences in the PorV complex. **b**, Locations of the cap and step domains in the PorV complex.

c, Insert sequence 1 is inserted within SprA β -strand 28. **d**, The PPI active site contains a proline-containing SprA surface loop (red).

the T9SS^{20–22}. A lateral opening in the wall of the SprA barrel links the barrel interior to the extracellular membrane surface (Fig. 1g, Extended Data Fig. 3b, Supplementary Videos 3, 4). In the PorV complex the lateral opening is completely blocked by PorV, which binds across its exterior face (Fig. 1g). In the Plug complex, access to the lateral opening is unobstructed, but the periplasmic end of the barrel is occluded by the Plug protein. Thus, the SprA cavity is opened to opposite sides of the membrane in the two complexes and neither complex provides an unimpeded pathway across the outer membrane. The size of the lateral opening is larger (about 45 Å diameter) in the Plug complex than the PorV complex owing to increased disorder in the loops surrounding the opening in the absence of bound PorV (Extended Data Fig. 3b).

The 36-strand barrel of SprA is the largest single polypeptide transmembrane β -barrel identified to date¹⁹, far exceeding the size of the 26-strand lipopolysaccharide transporter LptD (Fig. 3d). Indeed, the LptD protein has an internal pore that is 2.5 times smaller in cross-sectional area than that of SprA^{23,24}. The secretin family of outer membrane proteins form a channel of a similar size to SprA, but are constructed from symmetric, higher-order, oligomers of small subunits rather than a single asymmetric polypeptide. For example, the type II secretion system secretin is a 60-stranded barrel built from 15 copies of an approximately 40-kDa GspD subunit²⁵ (Fig. 3d). Notably, the outer membrane curli secretion channel, which is formed from nine CsgG proteins, has the same number of β -strands as SprA but a substantially narrower pore²⁶ (Fig. 3d).

Single polypeptide outer membrane β -barrels are inserted into the membrane by the Bam machinery²⁷. If SprA also uses this machinery,

it would be the only known Bam-integrated β -barrel that possesses large folded extracellular domains. The size and complexity of the SprA fold makes it difficult to predict how insertion of the barrel and folding of the extracellular domains are coordinated, although it is possible that the folded domains are individually exported through the SprA pore once assembly of the barrel is completed.

Structural analysis of SprA partner proteins

Of the three proteins that co-purify with SprA, PPI is the only one that is present in both the PorV and Plug complex structures (Fig. 1g). Unexpectedly for a protein with peptidyl-prolyl *cis-trans* isomerase activity, the PPI protein is located on the extracellular side of the membrane. PPI contacts SprA through two insertions near the C terminus of the SprA barrel (Fig. 2a). These insertions overhang the barrel edge and form fingers that ‘grab’ PPI (Figs. 1g, 2a). It is notable that one of the insertions occurs in the middle of a barrel strand (β -strand 28), splitting the strand in two and burying the insertion junction within the membrane bilayer (Fig. 2c). To the best of our knowledge, this is the first time that insertion within a strand has been observed in an outer membrane β -barrel protein. Binding of PPI to the overhang domain is mediated, in part, by insertion of a Pro-containing SprA surface loop into the PPI active site²⁸ (Fig. 2d). The lipidated N terminus of PPI is not visible in the EM density. However, the PPI molecule is appropriately oriented to allow membrane insertion of the lipid anchor.

In the PorV complex, the PorV protein binds across the lateral opening of SprA on the opposite side of the molecule to PPI (Fig. 1g). The packing interactions between PorV and the surface of SprA around the

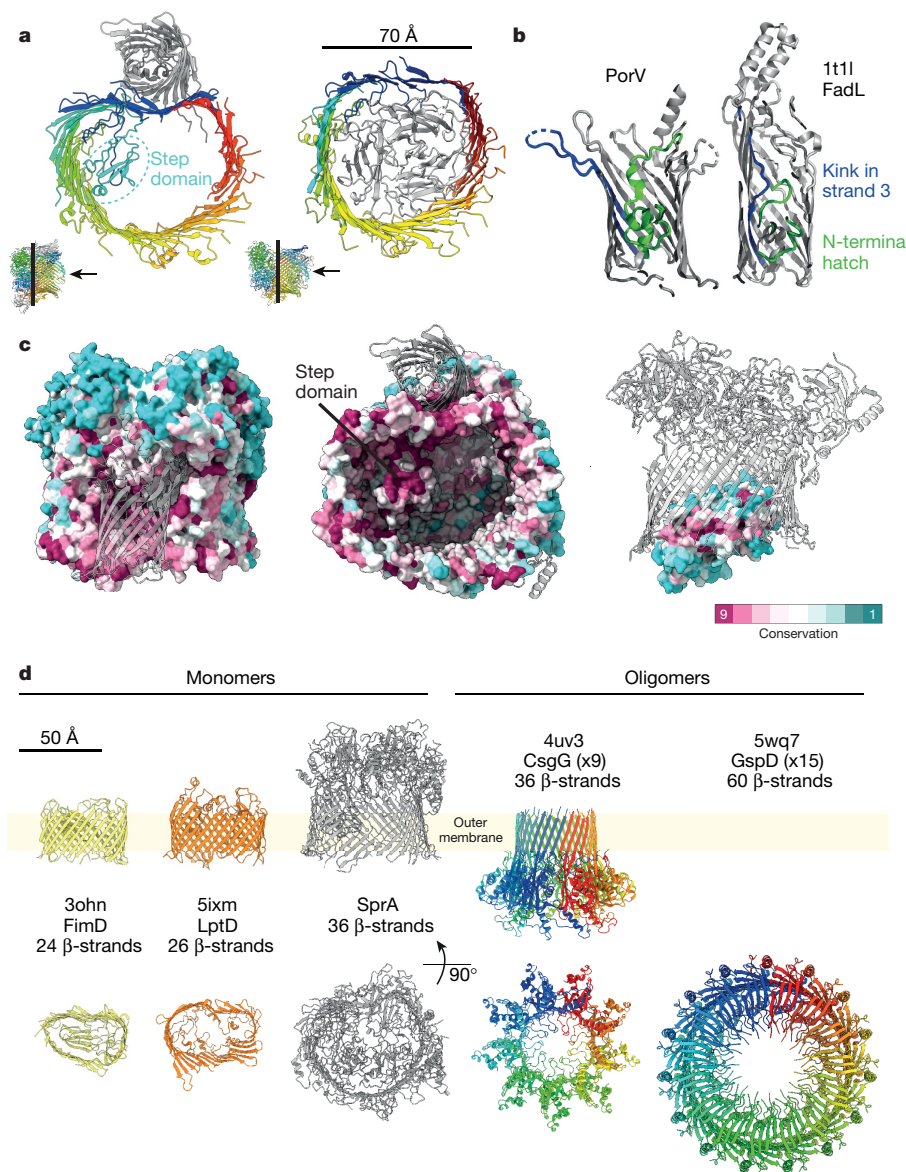


Fig. 3 | Structural analysis of the SprA translocon complexes. **a**, SprA barrel morphologies in the PorV (left) and Plug complexes (right) viewed from the periplasm and with cap domains cut away for clarity (insert). Grey, PorV or Plug. **b**, Structural comparison of PorV with FadL. **c**, Surface conservation of SprA (left and centre, shown in the context of the PorV

complex) and Plug (right, shown in the Plug complex) as determined by ConSurf³¹. **d**, Structural comparison of SprA with other large β -barrel bacterial outer membrane proteins (PDB accession codes given above protein names).

lateral opening are extensive and conserved (Fig. 3c). PorV is a member of the FadL family of 14-strand β -barrel outer membrane porins²⁹ (Fig. 3b, Extended Data Fig. 3a). As in other FadL family members, the pore of PorV is plugged by the N-terminal region of the protein (Fig. 3b). In previously characterized FadL family members, barrel strand 3 kinks inwards to sit over an N-terminal 'hatch' domain located within the periplasmic end of the barrel pore. However, in PorV, the N-terminal region of the protein fills the pore and strands 3 and 4 bend outwards from the PorV barrel axis with their inter-strand loop penetrating the interior of the SprA barrel through the lateral opening (Fig. 1g, Supplementary Video 3). The PorV barrel is tilted by 25° relative to the SprA barrel (Extended Data Fig. 3c). From the positions of the belts of aromatic residues on the surface of PorV that would normally lie at the membrane polar–apolar interfaces (Extended Data Fig. 3c), we deduce that PorV adopts a more vertical position in the membrane bilayer when not in complex with SprA.

In the Plug complex structure, the disc-shaped Plug protein is inserted into the periplasmic end of the SprA barrel to form a tight

seal between the barrel wall and the Plug rim (Fig. 1g, Supplementary Video 4). The interaction between SprA and the Plug is mediated solely through the Plug rim. Surface sequence conservation on the Plug is restricted to this contact interface (Fig. 3c), which suggests that the Plug has no other functional interactions. The location of the Plug protein in the Plug complex of SprA is sterically incompatible with the position of the Step domain seen in the PorV complex, explaining why the EM density for the Step domain is relocated and disordered in the Plug complex relative to the PorV complex.

A comparison of the two SprA complex structures shows that the cross-sectional shape of the SprA barrel changes from kidney-like in the PorV complex to almost circular in the Plug complex (Fig. 3a, Extended Data Fig. 3b, Supplementary Video 5). The Plug protein has perfect shape complementarity with the circular conformer of the barrel but would be unable to fit within the narrower kidney-shaped conformer. By contrast, the binding site for PorV is formed by the concave exterior surface of the kidney-shaped conformer of the SprA barrel, which

does not exist in the circular conformer. Thus, in each of the two SprA complex structures, the conformation of the barrel is intimately linked to the interaction that SprA makes with its partner protein, and binding of PorV and the Plug protein to SprA is mutually exclusive. There are no significant changes in the folded extracellular domains or PPI between the PorV and Plug complexes (Supplementary Video 5). Consequently, our structures provide no evidence for re-organization of the extracellular domains of SprA during the transport process.

SprA partner protein function

The identification of the PPI and Plug proteins in the T9SS translocon complexes was unanticipated. We examined the role of these proteins in T9SS by constructing corresponding *F. johnsoniae* deletion strains and comparing the effect on T9SS function with strains lacking either SprA or PorV. SprA was stably expressed in cells lacking any of the other three proteins found in the SprA complexes (Extended Data Fig. 4a). Thus, these three partner proteins are not required for SprA biogenesis. T9SS function was assessed by analysing the secreted proteome and by examining gliding motility, which depends on T9SS-exported adhesins^{8,9} (Fig. 4a–c, Extended Data Fig. 4b–d, Supplementary Video 1). In agreement with a previous report⁹, we observed that the loss of PorV prevents the secretion of most T9SS substrates but still permits gliding motility. By contrast, neither loss of the Plug protein nor loss of the PPI subunit had a detectable effect on T9SS function. Thus, the T9SS does not require the Plug or PPI protein for the transport process. Instead, a consideration of our structural data suggests that the role of the Plug protein might be to prevent non-specific leakage of periplasmic contents through the SprA channel. We reasoned that this function of the Plug protein would be revealed in a strain lacking both PorV and Plug because this strain would not be able to control movement of molecules through SprA. In agreement with this expectation, we found that a *porV plug* double mutant became sensitive to vancomycin (Fig. 4d, Extended Data Fig. 4e), an antibiotic that is normally excluded from Gram-negative bacteria because it is too large to fit through the protein channels in the outer membrane³⁰. Notably, vancomycin resistance was restored if *sprA* was also deleted, showing that SprA mediates the increased outer membrane permeability observed in the absence of PorV and Plug. Control experiments confirmed that the *porV plug* double mutant retained native levels of the SprA protein and phenocopied the transport phenotype of a *porV* single mutant (Fig. 4a–c, Extended Data Fig. 4a–d). Removal of the Plug protein alone also led to vancomycin sensitivity (Fig. 4d, Extended Data Fig. 4e), suggesting that the extracellular end of the SprA pore is routinely opened during the normal operation of the translocon. The results of our vancomycin sensitivity experiments show that the SprA protein forms a large outer membrane channel and that the Plug protein has a role in preventing leakage of small molecules through SprA.

Discussion

Our structural data show that SprA forms a water-filled conduit across the outer membrane that is large enough to allow the passage of folded proteins. The lateral position of the extracellular exit from this channel ensures that exported T9SS substrates are directed towards the membrane surface for subsequent modification by the attachment complex. Our structures show that the two ends of the transport pathway through SprA are sealed by the Plug protein and PorV, but that binding of these partner proteins is mutually exclusive. These observations suggest a model for the T9SS translocon mechanism in which the two ends of the transport channel are alternately gated (Fig. 4e). In this model, the PorV-containing state of the translocon allows uptake of substrates from the periplasm into the SprA channel, whereas the Plug complex represents the translocon after substrate release. As PorV is known to interact with T9SS substrate proteins¹³, we hypothesize that the translocon recognizes substrate proteins in the interior of the SprA barrel through interactions between the substrate CTD and the portions of PorV that are accessible through the lateral opening. This interaction is proposed to trigger the release of the PorV-substrate complex from

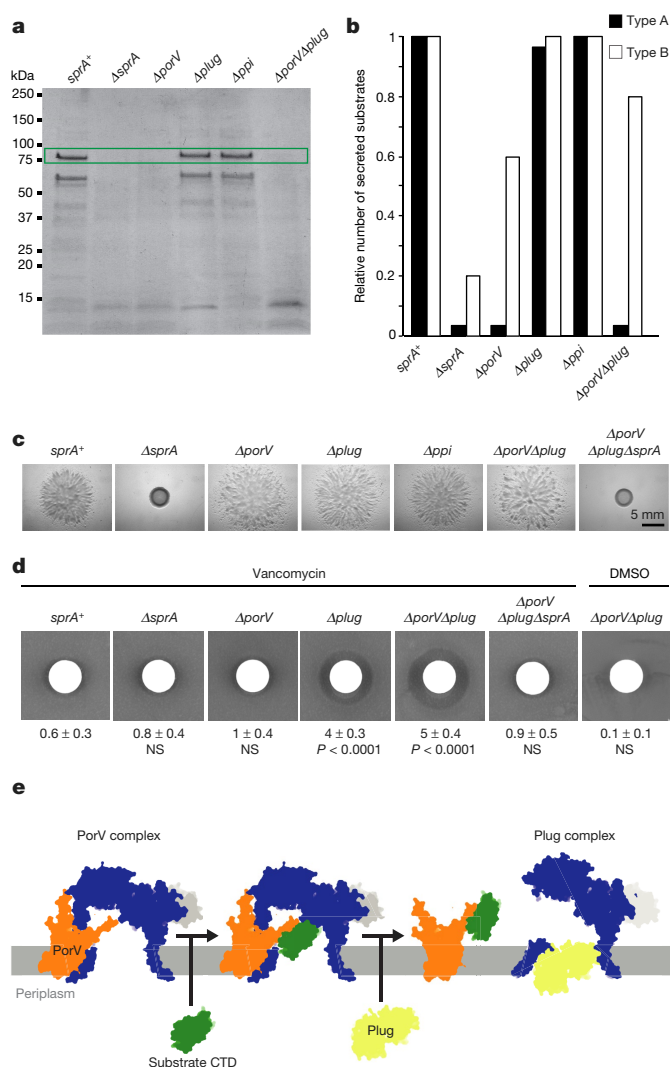


Fig. 4 | Biological consequences of removing SprA partner proteins.

a, b, Secretome analysis of culture supernatants. **a**, Coomassie-stained SDS-PAGE gel. The major T9SS-secreted chitinase ChiA is outlined in green. Similar data were obtained for three biological repeats. For gel source data, see Supplementary Fig. 1. **b**, T9SS substrates detected by proteomics at more than 1% of the protein abundance in strain *sprA*⁺, for which 30 type A (TIGR04183 family) and 5 type B (TIGR04131 family) CTD-dependent proteins were detected. **c**, Spreading (gliding) morphology of colonies on agar. Similar data were obtained for three biological repeats. **d**, Vancomycin sensitivity by disc diffusion assay. Mean \pm s.d. radius of inhibition (in mm) measured from the disc edge ($n = 4$ for Δ *porV* and Δ *plug*; $n = 5$ otherwise); statistical significance calculated by one-way ANOVA with post-hoc Dunnett's test using *sprA*⁺ as control group. NS, not significant. **a–d**, *sprA*⁺ strains express the Twin-Strep–SprA fusion. **e**, Model for the T9SS translocon mechanism.

the translocon. Once the periplasmic end of the SprA pore is clear of substrate protein, we suggest that the Plug protein seals the SprA channel until a PorV molecule is once more bound at the lateral opening. This mechanistic model leads to unidirectional transport of the substrate through the T9SS translocon.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0693-y>.

Received: 24 July 2018; Accepted: 13 September 2018;

Published online 7 November 2018.

1. Veith, P. D., Glew, M. D., Gorasia, D. G. & Reynolds, E. C. Type IX secretion: the generation of bacterial cell surface coatings involved in virulence, gliding motility and the degradation of complex biopolymers. *Mol. Microbiol.* **106**, 35–53 (2017).
2. Lasica, A. M., Ksiazek, M., Madej, M. & Potempa, J. The type IX secretion system (T9SS): highlights and recent insights into its structure and function. *Front. Cell. Infect. Microbiol.* **7**, 215 (2017).
3. Li, N. et al. The type IX secretion system is required for virulence of the fish pathogen *Flavobacterium columnare*. *Appl. Environ. Microbiol.* **83**, e01769-17 (2017).
4. Pérez-Pascual, D. et al. More than gliding: involvement of GldD and GldG in the virulence of *Flavobacterium psychrophilum*. *Front. Microbiol.* **8**, 2168 (2017).
5. Shrivastava, A., Roland, T. & Berg, H. C. The screw-like movement of a gliding bacterium is powered by spiral motion of cell-surface adhesins. *Biophys. J.* **111**, 1008–1013 (2016).
6. McBride, M. J. & Nakane, D. *Flavobacterium* gliding motility and the type IX secretion system. *Curr. Opin. Microbiol.* **28**, 72–77 (2015).
7. Shoji, M. et al. Por secretion system-dependent secretion and glycosylation of *Porphyromonas gingivalis* hemin-binding protein 35. *PLoS One* **6**, e21372 (2011).
8. Kulkarni, S. S., Zhu, Y., Brendel, C. J. & McBride, M. J. Diverse C-terminal sequences involved in *Flavobacterium johnsoniae* protein secretion. *J. Bacteriol.* **199**, e00884-16 (2017).
9. Kharade, S. S. & McBride, M. J. *Flavobacterium johnsoniae* PorV is required for secretion of a subset of proteins targeted to the type IX secretion system. *J. Bacteriol.* **197**, 147–158 (2015).
10. Heath, J. E. et al. PG1058 is a novel multidomain protein component of the bacterial type IX secretion system. *PLoS One* **11**, e0164313 (2016).
11. Gorasia, D. G. et al. Structural insights into the PorK and PorN components of the *Porphyromonas gingivalis* type IX secretion system. *PLoS Pathog.* **12**, e1005820 (2016).
12. Leone, P. et al. Type IX secretion system PorM and gliding machinery GldM form arches spanning the periplasmic space. *Nat. Commun.* **9**, 429 (2018).
13. Glew, M. D. et al. PorV is an outer membrane shuttle protein for the type IX secretion system. *Sci. Rep.* **7**, 8790 (2017).
14. Glew, M. D. et al. PG0026 is the C-terminal signal peptidase of a novel secretion system of *Porphyromonas gingivalis*. *J. Biol. Chem.* **287**, 24605–24617 (2012).
15. Gorasia, D. G. et al. *Porphyromonas gingivalis* type IX secretion substrates are cleaved and modified by a sortase-like mechanism. *PLoS Pathog.* **11**, e1005152 (2015).
16. Nelson, S. S., Glocka, P. P., Agarwal, S., Grimm, D. P. & McBride, M. J. *Flavobacterium johnsoniae* SprA is a cell surface protein involved in gliding motility. *J. Bacteriol.* **189**, 7145–7150 (2007).
17. Saiki, K. & Konishi, K. Identification of a *Porphyromonas gingivalis* novel protein Sov required for the secretion of gingipains. *Microbiol. Immunol.* **51**, 483–491 (2007).
18. McBride, M. J. & Zhu, Y. Gliding motility and Por secretion system genes are widespread among members of the phylum *Bacteroidetes*. *J. Bacteriol.* **195**, 270–278 (2013).
19. Schiffrin, B., Brockwell, D. J. & Radford, S. E. Outer membrane protein folding from an energy landscape perspective. *BMC Biol.* **15**, 123 (2017).
20. Lasica, A. M. et al. Structural and functional probing of PorZ, an essential bacterial surface component of the type-IX secretion system of human oral-microbiome *Porphyromonas gingivalis*. *Sci. Rep.* **6**, 37708 (2016).
21. Goulas, T. et al. Structure and mechanism of a bacterial host-protein citrullinating virulence factor, *Porphyromonas gingivalis* peptidylarginine deiminase. *Sci. Rep.* **5**, 11969 (2015).
22. de Diego, I. et al. *Porphyromonas gingivalis* virulence factor gingipain RgpB shows a unique zymogenic mechanism for cysteine peptidases. *J. Biol. Chem.* **288**, 14287–14296 (2013).
23. Qiao, S., Luo, Q., Zhao, Y., Zhang, X. C. & Huang, Y. Structural basis for lipopolysaccharide insertion in the bacterial outer membrane. *Nature* **511**, 108–111 (2014).
24. Dong, H. et al. Structural basis for outer membrane lipopolysaccharide insertion. *Nature* **511**, 52–56 (2014).
25. Yan, Z., Yin, M., Xu, D., Zhu, Y. & Li, X. Structural insights into the secretin translocation channel in the type II secretion system. *Nat. Struct. Mol. Biol.* **24**, 177–183 (2017).
26. Goyal, P. et al. Structural and mechanistic insights into the bacterial amyloid secretion channel CsgG. *Nature* **516**, 250–253 (2014).
27. Noinaj, N., Gumbart, J. C. & Buchanan, S. K. The β -barrel assembly machinery in motion. *Nat. Rev. Microbiol.* **15**, 197–204 (2017).
28. Quistgaard, E. M. et al. Molecular insights into substrate recognition and catalytic mechanism of the chaperone and FKBP peptidyl-prolyl isomerase SlyD. *BMC Biol.* **14**, 82 (2016).
29. van den Berg, B., Black, P. N., Clemons, W. M., Jr & Rapoport, T. A. Crystal structure of the long-chain fatty acid transporter FadL. *Science* **304**, 1506–1509 (2004).
30. Zgurskaya, H. I., López, C. A. & Gnanakaran, S. Permeability barrier of Gram-negative cell envelopes and approaches to bypass it. *ACS Infect. Dis.* **1**, 512–522 (2015).
31. Ashkenazy, H. et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* **44**, W344–W350 (2016).

Acknowledgements We thank M. McBride and Y. Zhu for providing reagents for the genetic manipulation of *F. johnsoniae*; A. Shrivastava and H. Berg for advice on measuring gliding motility; L. Lavis for supplying the Janelia Fluor 646 HaloTag ligand; S. Hickman for advice on fluorescence imaging; and O. Meacock and K. Foster for providing additional imaging facilities. We acknowledge the use of Central Oxford Structural Microscopy and Imaging Centre (COSMIC), the Oxford Micron Advanced Imaging Facility, and the Oxford Advanced Proteomics Facility. This work was supported by Wellcome Trust Investigator Awards 107929/Z/15/Z and 100298/Z/12/Z. COSMIC was supported by a Wellcome Trust Collaborative Award 201536/Z/16/Z, the Wolfson Foundation, a Royal Society Wolfson Refurbishment Grant, the John Fell Fund, and the EPA and Cephalosporin Trusts.

Reviewer information Nature thanks M. McBride and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions F.L. carried out all genetic and biochemical work. J.C.D. collected EM data. J.C.D. and S.M.L. determined the structure. B.C.B. conceived the project. All authors interpreted data and wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0693-y>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0693-y>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to S.M.L. or B.C.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Bacterial strains and growth conditions. All strains and plasmids used in this work are listed in Extended Data Table 2. *F. johnsoniae* was routinely grown in Casitone Yeast Extract (CYE) medium³² at 25 °C with shaking. For some studies the cells were cultured on Motility medium³³ or PY2 medium³⁴ as described below. Where required, 100 µg/ml erythromycin was included in the growth medium.

Genetic constructs. A suicide plasmid to introduce an in-frame unmarked deletion of *sprA* was produced as follows. A 2.5-kb fragment corresponding to the first 57 bp of *sprA* together with the region directly upstream was amplified using primers FL199 and FL200. Similarly, a 2.5-kb fragment corresponding to the last 72 bp of *sprA* and the region directly downstream was amplified with primers FL201 and FL202. Finally, plasmid pYT313³⁵ was linearized by PCR amplification using FL203 and FL204. The three amplicons were then assembled by Gibson cloning resulting in plasmid pFL58. Suicide plasmids to construct in-frame deletions in *porV* (*fjoh_1555*), *plug* (*fjoh_1759*) and *ppi* (*fjoh_4997*) were constructed in an analogous way using the oligonucleotides specified in Supplementary Table 1.

To produce a strain in which a HaloTag was inserted between the signal peptide and the mature region of SprA, we constructed the following suicide plasmid. A 2.5-kb fragment corresponding to the first 63 bp of *sprA* and the region directly upstream was amplified using primers FL217 and FL218. This fragment was inserted between the NcoI and SpeI sites of pGEM-T Easy to generate pFL56. A 2.7-kbp fragment covering nucleotides 55–2770 of *sprA* was amplified using primers FL219 and FL220. This fragment was inserted between the SpeI and SacI sites of pFL56 to generate pFL57. The HaloTag-coding sequence was amplified from plasmid pHTC HaloTag CMV-neo using primers FL178 and FL227 and the resulting fragment was ligated between the BamHI and SpeI sites of pFL57, yielding pFL61. The *halotag::sprA* fusion sequence from pFL61 was then introduced into pYT313 using Gibson assembly to generate pFL64. The suicide plasmid used to produce a strain with a Twin-Strep tag between the signal peptide and the mature region of SprA was constructed by replacing the HaloTag-coding sequence in pFL61 by a Twin-Strep tag-coding sequence using Q5 site-directed mutagenesis (New England Biolabs) with primers FL221 and FL222, yielding pFL66. The resulting fusion sequence was then introduced into pYT313 using Gibson assembly to generate pFL67.

Suicide plasmids were introduced into the appropriate *F. johnsoniae* background strain by biparental mating using *E. coli* S17-1³⁶ as donor strain, as previously described³², and erythromycin resistance was used to select cells with chromosomally integrated plasmid. One of the resulting clones was grown overnight in CYE without antibiotics to allow for loss of the plasmid backbone and then plated onto CYE agar containing 5% sucrose. Sucrose-resistant colonies were screened by PCR for the presence of the desired chromosomal modification and then verified by sequencing.

Purification of SprA complexes. *F. johnsoniae* FL_012 was cultured aerobically at 25 °C for 22 h in 10 l of CYE medium. Cells were harvested by centrifugation at 6,000g for 25 min and stored at –20 °C until further use. All purification steps were carried out at 4 °C. Cell pellets were resuspended in buffer W (100 mM Tris-HCl, pH 8, 150 mM NaCl, 1 mM EDTA) containing 30 µg/ml DNase I, 400 µg/ml lysozyme and 1 mM PMSEF at a ratio of 7.5 ml of buffer to 1 g of cell pellet. Cells were incubated on ice for 30 min before being lysed by two passages through a TS series 1.1 kW cell disruptor (Constant System Ltd) at 30,000 PSI. Unbroken cells were removed by centrifugation at 10,000g for 10 min. The supernatant was recovered and total membranes were collected by centrifugation at 150,000g for 75 min. Membranes were resuspended in buffer W to a protein concentration of 6.5 mg/ml and solubilized by incubation with 1% (w/v) lauryl maltose neopentyl glycol (LMNG) (Anatrace) for 2 h. Insoluble material was removed by centrifugation at 150,000g for 60 min. The supernatant was then circulated through a StrepTrap HP column (GE Healthcare) overnight. The column was washed with 15 column volumes (CV) of buffer W containing 0.01% LMNG (buffer WD) and bound proteins were eluted with 6 CV buffer WD containing 2.5 mM desthiobiotin. The eluate was concentrated to 500 µl using a 100-kDa molecular weight cutoff (MWCO) Amicon ultra-15 centrifugal filter unit, and injected onto a Superose 6 Increase 10/300 GL column (GE Healthcare) previously equilibrated in buffer WD. Peak fractions were collected and concentrated using a 100-kDa MWCO Vivaspin 500 column.

Cryo-EM sample preparation and data collection. The samples imaged were purified SprA at $A_{280\text{nm}} = 1.0$ in buffer WD, purified SprA at $A_{280\text{nm}} = 2.0$ in buffer WD with 1.5 mM fluorinated octyl-choline 8 (Anatrace), or purified SprA at $A_{280\text{nm}} = 2.3$ in buffer WD with 0.7 mM fluorinated octyl maltoside (Anatrace). Four microlitres of each sample was applied onto glow-discharged holey carbon coated grids (Quantifoil 300 mesh, Au R1.2/1.3), adsorbed for 10 s, blotted for 2 s at 100% humidity at 4 °C and plunge frozen in liquid ethane using a Vitrobot Mark IV (FEI).

Data were collected in counting mode on a Titan Krios G3 (FEI) operating at 300 kV with a GIF energy filter (Gatan) and K2 Summit detector (Gatan). Data sets were collected for each purified sample at a sampling of 0.85 Å/pixel, 6.5 e[–]/Å²/s, 8 s exposure, total dose 52 e[–]/Å², 20 fractions written totalling 14,553 movies. Motion correction and dose-weighting were performed with SIMPLE-unblur³⁷ and contrast transfer functions (CTFs) of the summed micrographs were calculated using CTFFIND4³⁸. Dose-weighted micrographs were subjected to picking using SIMPLE³⁷ and extraction with a 300 × 300 Å box in Relion 2.0³⁹, which was used for all further processing. Reference-free 2D classification was performed separately for each data set. A low-resolution initial model generated from the fluorinated octyl maltoside data set was used as a reference for 3D classification of all particles. Particles associated with PorV complex or Plug complex models were subjected to auto-refinement using a corresponding lowpass-filtered map. A final masked auto-refinement was carried out with the resulting models as references using 240,826 particles for the PorV complex or 118,090 particles for the Plug complex. Post-processing was carried using a soft mask with *B*-factor of –155 Å² (PorV complex) or –157 Å² (Plug complex) and a calibrated pixel size of 0.82 Å/pixel. Gold standard Fourier shell correlations using the 0.143 criterion led to resolution estimates of 3.5 Å for the PorV complex and 3.7 Å for the Plug complex. Local resolution estimations were calculated within Relion 2.0.

The EM processing workflow is detailed schematically in Extended Data Fig. 5. **Model building and refinement.** A backbone trace of the PorV complex was manually made using program COOT³⁸ in a 4.2 Å map. 2,423 α-carbons were placed. This is more than the full length of SprA, implying that other components of the system had been co-purified and imaged with the tagged SprA. Inspection of this model allowed us to identify an unconnected domain at the periphery of the complex. A structure-based search of the PDB using PDBFold revealed this domain to be of the FKBP-prolyl peptide isomerase family of proteins (chain A of PDB 5hua being the highest hit). As a member of this family had been identified in the sample by proteomics, this portion of the trace was assumed to be Fjoh_4997 and sequence for this protein was docked using 5hua as a guide. A similar search with the small beta barrel domain established OmpG (PDB 4ctd) as the closest structural homologue and confirmed assignment of N- and C-terminal strands within the barrel. Incorporation of additional particles and further refinement of the volume yielded a 3.5 Å map with clear sidechain densities throughout the volume. The sidechains decorating the small barrel structure were consistent with the sequence of the PorV homologue (Fjoh_1555) identified in the sample by proteomics.

The Plug complex was built using the PorV complex (with the PorV barrel removed) as a starting model, which left a large ordered portion of density unoccupied. As for PPI and PorV components, an α-carbon trace was built into this density and PDBFold identified a protein of unknown function (PDB 4r7f, a hypothetical protein from *Parabacteroides merdae*) as sharing the same fold.

Multiple rounds of rebuilding (in both the globally sharpened and model-based sharpened maps) and real-space refinement in Phenix⁴⁰ using secondary structure, rotamer and Ramachandran restraints yielded the final models described in Extended Data Table 1. The PorV complex was used as a reference model to provide additional restraints for refinement of the Plug complex.

Amino acid residues are numbered relative to the start of the corresponding native precursor protein. See Extended Data Table 1 for cryo-EM data collection, refinement and validation statistics. Protein structure figures were prepared using Pymol Version 2.0 (Schrödinger, LLC) and ChimeraX⁴¹.

Immunoblotting. For whole cell immunoblots, strains were cultured to mid-log phase ($OD_{600} = 0.3$ – 0.4) in motility medium. Proteins were then transferred onto polyvinylidene difluoride (PVDF) membranes and probed with primary antibodies (anti-StrepTag (34850, Qiagen), anti-HaloTag (G921A, Promega), or anti-ChiA) followed by appropriate secondary antibodies (anti-mouse (A4416, Merck) or anti-rabbit (31462, Pierce)). ChiA antibodies were raised against urea-solubilized inclusion bodies of a recombinant protein (expressed from plasmid pFL53) corresponding to the N-terminal glycoside hydrolase domain of ChiA.

Proteomics. Gel bands were cut and subjected to in-gel trypsin digestion as previously described⁴². The resulting digests were analysed by a nano-flow reversed-phase liquid chromatograph coupled to a Q Exactive Hybrid Quadrupole-Orbitrap mass spectrometer (Thermo Fisher Scientific). The digests were loaded onto a C18 PepMap100 pre-column (inner diameter 300 µm × 5 mm, 3 µm C18 beads; Thermo Fisher Scientific) and separated on a 50-cm reversed-phase C18 column (inner diameter 75 µm, 2 µm C18 beads) using a linear gradient from 10 to 35% of buffer B for 30 min at a flow rate of 200 nL/min (buffer A: 0.1% formic acid; buffer B: 0.1% formic acid in acetonitrile). All data were acquired in a data-dependent mode, automatically switching from mass spectrometry (MS) to collision-induced dissociation MS/MS on the top 10 most abundant ions with a precursor scan range of 350–2,000 *m/z*. MS spectra were acquired at a resolution of 70,000 and MS/MS scans at 17,500. Dynamic exclusion was enabled with an exclusion duration of 5 s and charge exclusion was applied to unassigned and mono-charged ions. Raw data files were processed for protein

identification using MaxQuant, version 1.5.0.35, integrated with the Andromeda search engine as described previously^{43–45}. The MS/MS spectra were searched against the *F. johnsoniae* UW101 Uniprot database. Precursor mass tolerance was set to 20 p.p.m. and MS/MS tolerance to 0.05 Da. Enzyme specificity was set to trypsin with a maximum of two missed cleavages. False discovery rate for protein and peptide spectral matches was set at 0.01.

To analyse the secreted proteome, strains were grown to late-log phase in CYE medium and a 10-ml sample of the culture collected when the OD₆₀₀ reached 4.5. Cells were removed by centrifugation at 9,000g for 25 min and the supernatant then filtered using a 0.22-µm syringe filter unit (Millipore). The supernatant was further clarified by centrifugation at 150,000g for 75 min. The resulting supernatant was concentrated to 350 µl using a 10-kDa molecular weight cutoff (MWCO) Amicon ultra-15 centrifugal filter unit. We subjected 25 µl of this secreted protein fraction to SDS-PAGE. Proteins were visualized by Coomassie blue staining and the whole gel lane was excised for mass spectrometry analysis.

Measurement of gliding motility on agar. Strains were grown overnight in PY2 medium, washed once with fresh medium, and resuspended in PY2 medium to OD₆₀₀ = 0.1. We then spotted 2 µl of cell suspension onto PY2 agar plates and incubated them at 25°C for 24 h. Colonies were imaged using a Zeiss AXIO Zoom.V16 microscope equipped with a Zeiss AxioCAM MRm CCD camera and Zeiss software (ZenPro 2012, version 1.1.1.0).

Microscopic observation of live cells. Cells for brightfield microscopic observation of gliding motility on glass were grown overnight in motility medium, inoculated into fresh medium at a 1:40 dilution, and grown for 5 h at 25°C and at 50 r.p.m. to a final OD₆₀₀ = 0.3–0.4. An aliquot of the cell culture was introduced into a tunnel slide, incubated for 5 min, washed twice with 100 µl motility medium and imaged.

For fluorescence imaging, cells were prepared in PY2 medium to reduce background fluorescence. Cells were cultured as above but in PY2 medium and to an OD₆₀₀ = 0.1–0.2. The HaloTag–SprA fusion was labelled by mixing 1 ml of this culture with 1 µl of a 10 µM stock solution of Janelia Fluor 646 HaloTag ligand⁴⁶ in dimethyl sulfoxide (DMSO) and incubating for a further 20 min at 25°C and 50 r.p.m. Cells were harvested by centrifugation, washed three times with PY2 medium, and resuspended in 50 µl of PY2 medium. To observe immobilized cells, 1 µl of cell suspension was spotted on a 1% agarose pad containing 50% PY2 medium. The spot was allowed to dry for 1 min and the cells were overlaid with a coverslip before imaging. To observe gliding motility on glass, an aliquot of the cell culture was introduced into a tunnel slide, incubated for 5 min, washed twice with 100 µl PY2 medium, and imaged.

All brightfield and fluorescence images were acquired at 25°C using a Nanoimager (Oxford Nanoimaging) equipped with a 640 nm 1W DPSS laser. Optical magnification was provided by a 100× oil-immersion objective (Olympus, numerical aperture (NA) 1.4) and images were acquired using an ORCA-Flash4.0 V3 CMOS camera (Hamamatsu). All fluorescence images were collected at 15% laser power. Measurement of the number of fluorescent SprA foci per cell was carried out using the ImageJ⁴⁷ plugin ThunderSTORM⁴⁸ with the following camera settings: pixel size, 117 nm; photoelectrons per A/D count, 0.46; base level, 100; and otherwise default settings. Photobleaching analysis of SprA foci was carried out using the Nanoimager software.

Vancomycin sensitivity assay. Strains were grown to mid-log phase (OD₆₀₀ = 0.3–0.4) in motility medium. Cultures were then diluted to OD₆₀₀ = 0.1 and 100 µl of the cell suspension used to inoculate CYE agar. Plates were allowed to dry for 20 min at room temperature. Filter paper discs containing 100 µg vancomycin hydrochloride (Merck) dissolved in DMSO were then placed on top of the agar. Inhibition zones were imaged after 48 h of incubation at 25°C using a G:box ChemiXX6

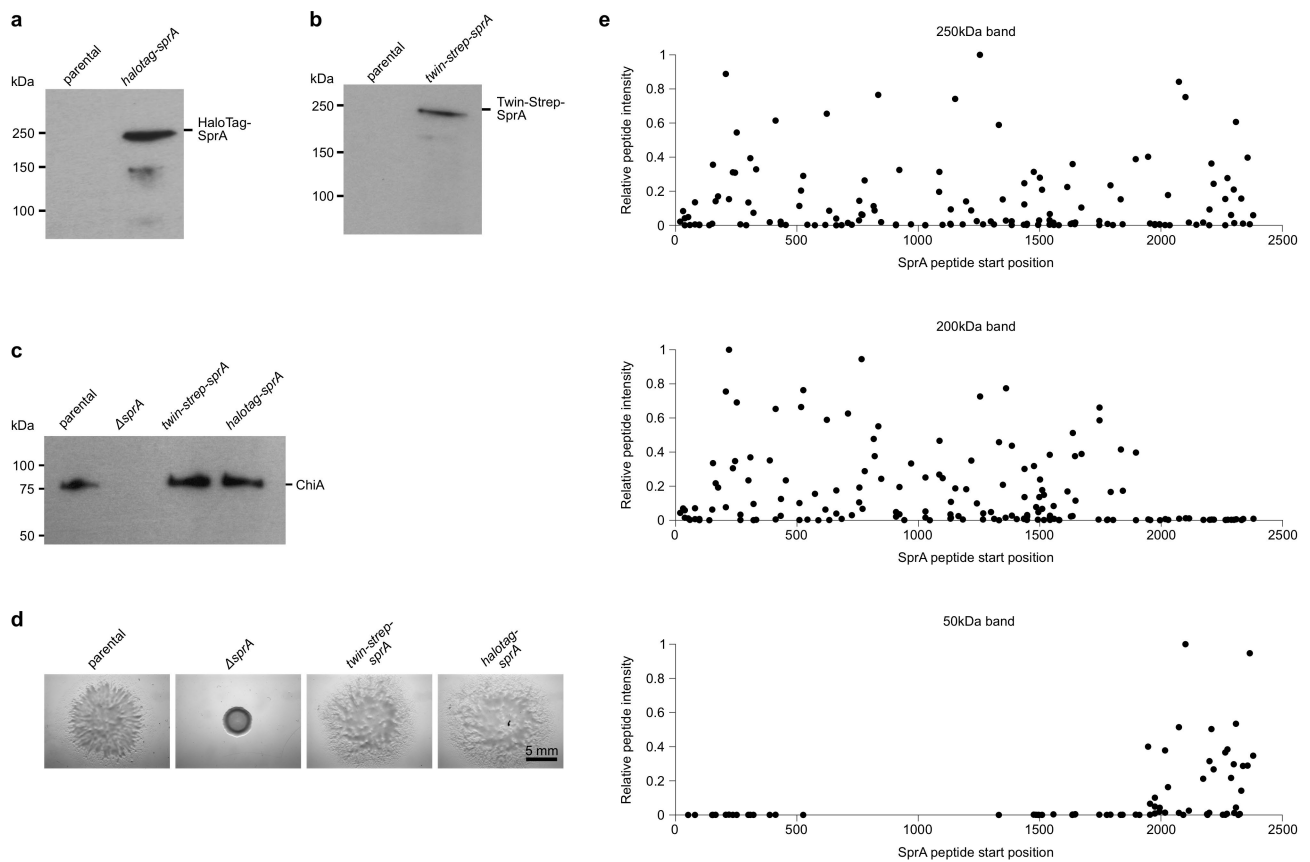
(Syngene) equipped with a SynGene 6MP camera and running GeneSys version 1.5.40 software.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

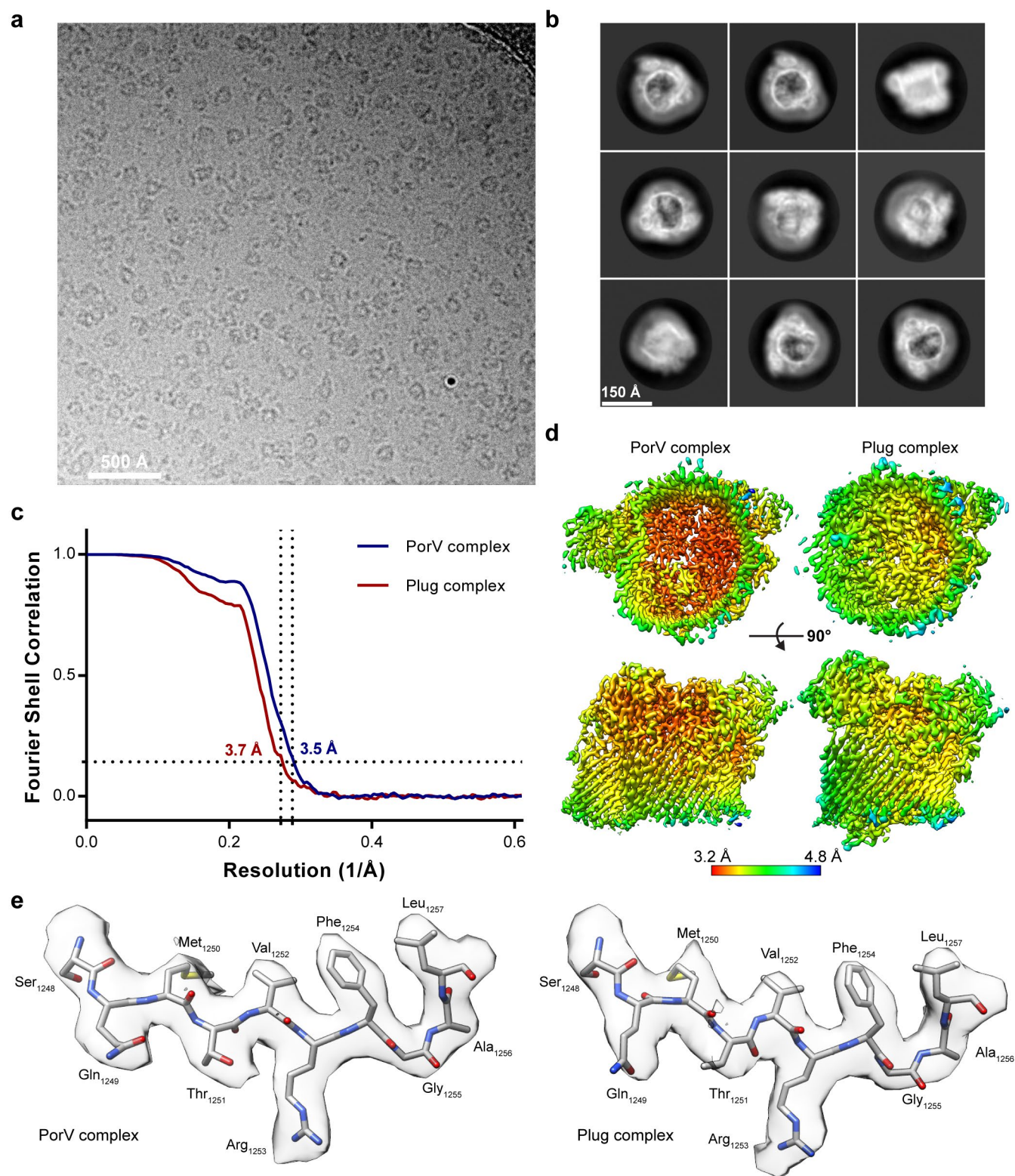
The cryo-EM volumes have been deposited in the Electron Microscopy Data Bank under accession codes EMD-0133 and EMD-0134, and the coordinates have been deposited in the Protein Data Bank under accession numbers 6h3i and 6h3j. Source Data for Figs. 1d and 4a, b are available with the online version of the paper.

- McBride, M. J. & Kempf, M. J. Development of techniques for the genetic manipulation of the gliding bacterium *Cytophaga johnsonae*. *J. Bacteriol.* **178**, 583–590 (1996).
- Liu, J., McBride, M. J. & Subramaniam, S. Cell surface filaments of the gliding bacterium *Flavobacterium johnsoniae* revealed by cryo-electron tomography. *J. Bacteriol.* **189**, 7503–7506 (2007).
- Agarwal, S., Hunnicutt, D. W. & McBride, M. J. Cloning and characterization of the *Flavobacterium johnsoniae* (*Cytophaga johnsonae*) gliding motility gene, *gldA*. *Proc. Natl Acad. Sci. USA* **94**, 12139–12144 (1997).
- Zhu, Y. et al. Genetic analyses unravel the crucial role of a horizontally acquired alginate lyase for brown algal biomass degradation by *Zobellia galactanivorans*. *Environ. Microbiol.* **19**, 2164–2181 (2017).
- Simon, R., Priefer, U. & Puhler, A. A broad host range mobilization system for in vivo genetic engineering—transposon mutagenesis in Gram-negative bacteria. *Bio/Technology* **1**, 784–791 (1983).
- Reboul, C. F., Eager, M., Elmlund, D. & Elmlund, H. Single-particle cryo-EM—Improved ab initio 3D reconstruction with SIMPLE/PRIME. *Protein Sci.* **27**, 51–61 (2018).
- Brown, A. et al. Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Crystallogr. D Biol. Crystallogr.* **71**, 136–153 (2015).
- Kimanius, D., Forsberg, B. O., Scheres, S. H. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *eLife* **5**, e18722 (2016).
- Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
- Goddard, T. D. et al. UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).
- Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V. & Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protocols* **1**, 2856–2860 (2006).
- Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
- Cox, J. et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
- Tyanova, S. et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
- Grimm, J. B., Brown, T. A., English, B. P., Lionnet, T. & Lavis, L. D. Synthesis of Janelia Fluor HaloTag and SNAP-Tag ligands and their use in cellular imaging experiments. *Methods Mol. Biol.* **1663**, 179–188 (2017).
- Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
- Ovesný, M., Křížek, P., Borkovec, J., Svindrych, Z. & Hagen, G. M. ThunderSTORM: a comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging. *Bioinformatics* **30**, 2389–2390 (2014).
- McBride, M. J. & Braun, T. F. GldI is a lipoprotein that is required for *Flavobacterium johnsoniae* gliding motility and chitin utilization. *J. Bacteriol.* **186**, 2295–2302 (2004).



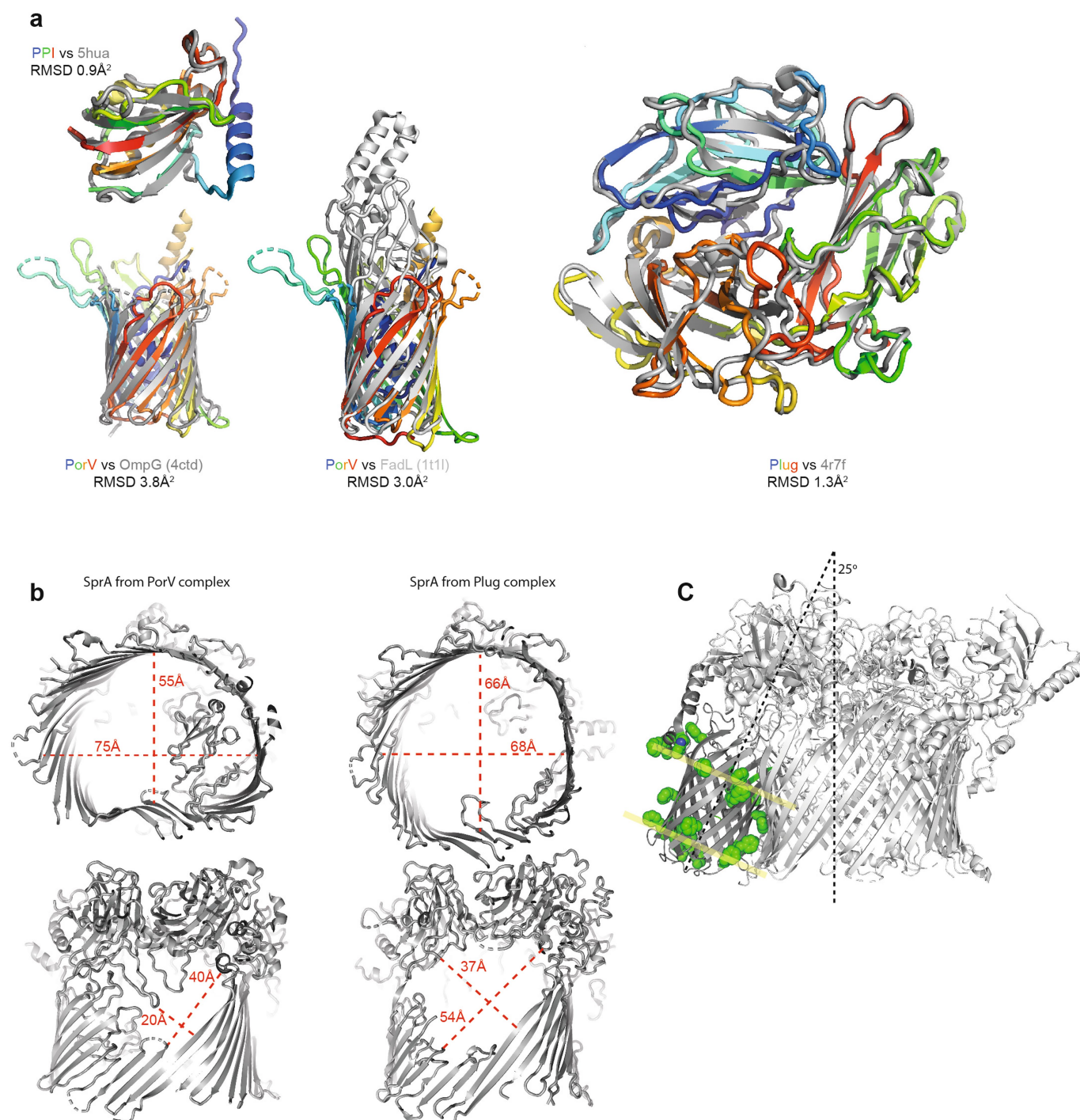
Extended Data Fig. 1 | Phenotypic analysis of strains expressing HaloTag and Twin-Strep SprA fusion proteins. a, b, Immunoblot analysis of HaloTag-SprA (a) and Twin-Strep-SprA (b) expression in whole-cell lysates. Similar data were obtained for three biological repeats. **c,** Immunoblot detection of levels of the T9SS-dependent chitinase ChiA in culture supernatants. Similar data were obtained for three biological repeats. **d,** T9SS-dependent spreading (gliding) morphology of colonies

on agar. Scale bar, 5 mm. Similar data were obtained for three biological repeats. **e,** Peptide mass spectrometry of the three highest molecular mass bands in Fig. 1d. Intensity values are normalized to the most abundant SprA peptide detected for each band. Peptide numbering is from the N terminus of the native SprA precursor sequence. For immunoblot source data, see Supplementary Fig. 1.



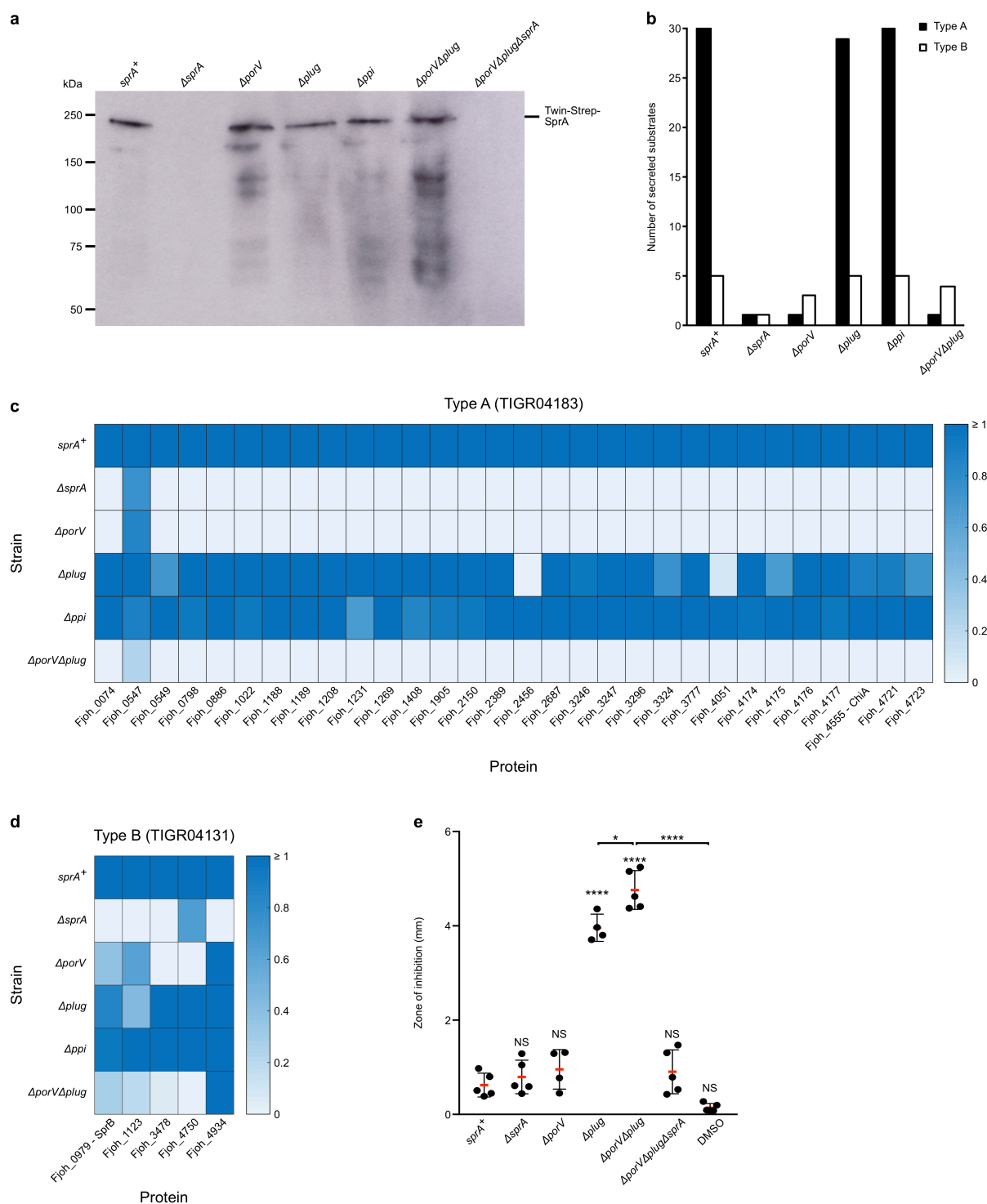
Extended Data Fig. 2 | Experimental quality and resolution estimation of SprA complexes. **a**, Representative micrograph of SprA complexes. **b**, Selected reference-free 2D class averages. **c**, Gold-standard FSC curves

of the final map calculated using a soft-edged mask. **d**, Local resolution estimates of the final maps. **e**, Representative density for SprA in the PorV complex (left) and Plug complex (right).



Extended Data Fig. 3 | Structural analysis of the SprA complex components. **a**, Structural alignment of SprA-bound proteins against the homology models used in initial sequence docking. **b**, Access routes to the SprA pore viewed from the periplasm (top) or towards the lateral opening (bottom). In the PorV complex two loops involved in coordinating PorV

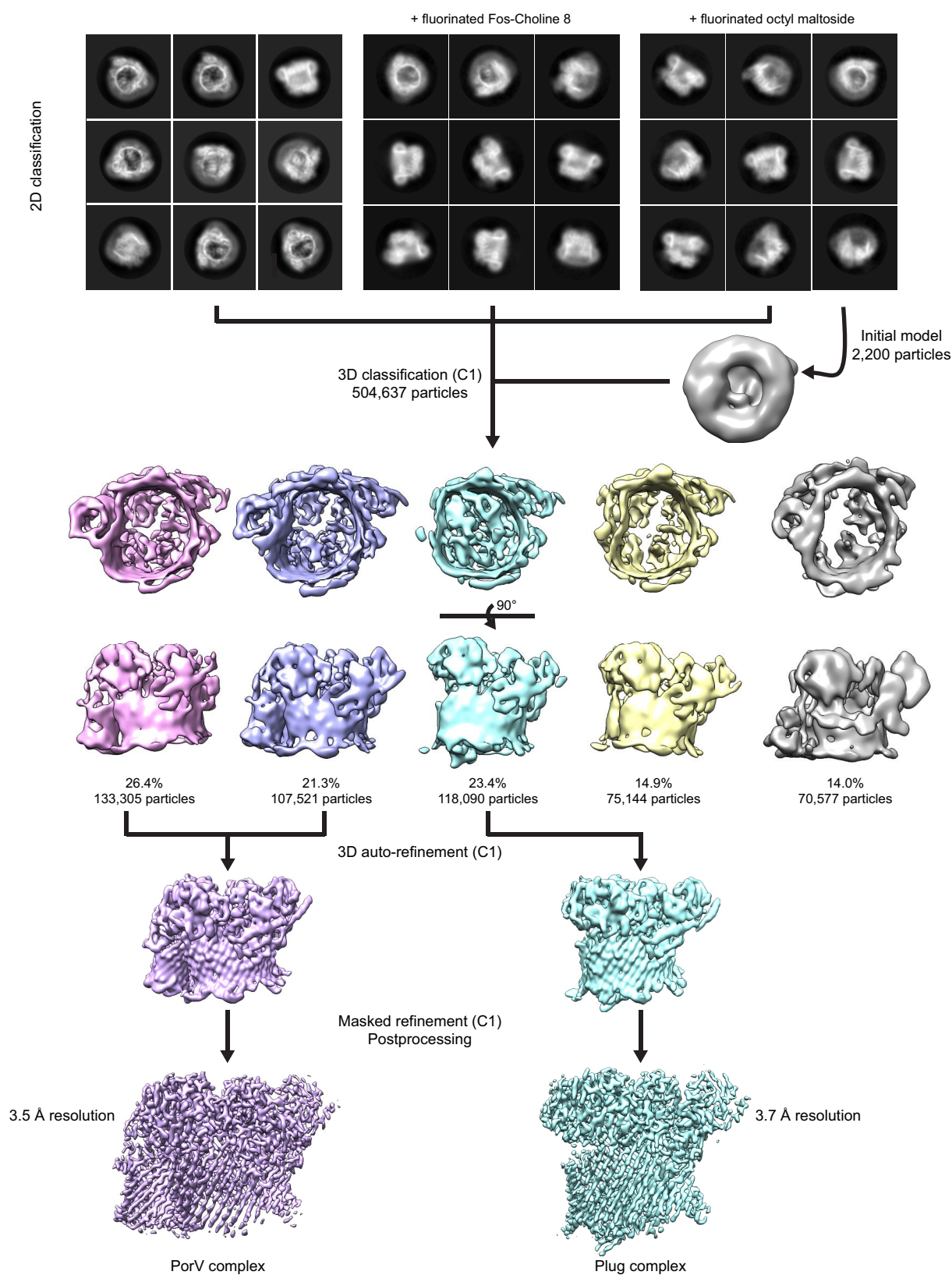
occlude the lateral opening. The step domain is poorly ordered in the Plug complex. For clarity, PorV, Plug, and PPI are not shown. **c**, The PorV barrel is tilted relative to the SprA barrel. Aromatic residues on the surface of PorV are shown in green spacefill.



Extended Data Fig. 4 | Phenotypes of *SprA* partner deletion strains.

a, Immunoblot analysis of Twin-Strep-SprA levels in whole-cell lysates. Similar data were obtained for three biological repeats. For immunoblot source data, see Supplementary Fig. 1. **b**, Quantification by liquid chromatography–mass spectrometry of T9SS substrates detected in cell culture supernatants according to CTD type. A detection threshold of more than 1% of the protein abundance in the *sprA*⁺ parental strain was applied. **c**, **d**, Heat map representations of secreted T9SS substrates from **b** with type A CTDs (**c**) or type B CTDs (**d**). Protein intensities for each protein are normalized to the level detected in the *sprA*⁺ parental strain.

e, Measurement of vancomycin inhibition zones in a disc diffusion assay. The mean radius of inhibition (red bar) was measured from the disc edge. Error bars represent s.d. ($n = 4$ for Δ *porV* and Δ *plug*; $n = 5$ for other strains) and statistical significance is shown above each measurement set from a one-way ANOVA with post-hoc Dunnett's test using *sprA*⁺ as control group (NS, not significant; **** $P < 0.0001$). Other comparisons (bracketed) use two-tailed unpaired *t*-tests; * $P = 0.0134$, **** $P < 0.0001$). A control for the DMSO solvent used to dissolve the vancomycin is shown. **a–e**, All *sprA*⁺ strains express the Twin-Strep-SprA fusion protein.



Extended Data Fig. 5 | Single-particle cryo-EM image processing workflow for the SprA complexes. Cryo-EM data sets for SprA complexes in LMNG, in the presence or absence of fluorinated detergents, were combined following 2D classification and subjected to 3D classification against a low-resolution model generated from the fluorinated octyl-maltoside

data set. Particle images corresponding to the PorV complex or the Plug complex were then independently refined. A soft mask, generated from these maps, was then used to perform masked refinement against the same particle images, resulting in global map resolutions of 3.5 Å for the PorV complex and 3.7 Å for the Plug complex.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	PorV complex (EMD-0133) (PDB 6h3i)	Plug complex (EMD-0134) (PDB 6h3j)
Data collection and processing		
Magnification	165,000	
Voltage (kV)	300	
Electron exposure (e-/Å ²)	52	
Defocus range (μm)	1.0-3.0	
Pixel size (Å)	0.82	
Symmetry imposed	C1	
Initial particle images (no.)	504,637	504,637
Final particle images (no.)	240,826	118,090
Map resolution (Å)	3.5	3.7
FSC threshold	0.143	0.143
Map resolution range (Å)	3.3-5.0	3.4-4.7
Refinement		
Initial model used (PDB code)	none	PorV-complex + 4r7f (chain C)
Model resolution (Å)	3.5	3.7
FSC threshold	0.143	0.143
Model resolution range (Å)	3.3-5.0	3.4-4.7
Map sharpening <i>B</i> factor (Å ²)	-155	-157
Model composition		
Non-hydrogen atoms	20,546	20,089
Protein residues	2,597 (SprA 2,124; FJOH_4997 128; FJOH_1555 345)	2,511 (SprA 1,988; FJOH_4997 126; FJOH_1749 397)
Ligands	0	0
<i>B</i> factors (Å ²)	50	58
Protein	(SprA 46; FJOH_4997 73; FJOH_1555 67)	(SprA 57; FJOH_4997 85; FJOH_1759 51)
R.m.s. deviations		
Bond lengths (Å)	0.009	0.007
Bond angles (°)	0.9	0.8
Validation		
MolProbity score	2.2	2.3
Clashscore	8.6	9.5
Poor rotamers (%)	0.4	0.05
Ramachandran plot		
Favored (%)	78	76
Allowed (%)	21	24
Disallowed (%)	0.2	0.04

PorV complex, model to map fit CC_mask = 0.84; Plug complex, model to map fit CC_mask = 0.83.

Extended Data Table 2 | Bacterial strains and plasmids used in this study

Strain	Genotype	Reference
<i>E. coli</i>		
S17-1	<i>pro</i> , <i>res^E</i> <i>hsdR17</i> (<i>rK^E</i> <i>mK[*]</i>) <i>recA^E</i> , <i>RP4-2-Tc::Mu-Km::Tn7</i> , <i>Tp^r</i>	36
<i>F. johnsoniae</i>		
UW101		49
FI_004	UW101 Δ <i>sprA</i>	This study
FI_012	UW101 <i>twin-strep-tag::sprA</i>	This study
FI_016	UW101 <i>halotag::sprA</i>	This study
FI_036	FI_012 Δ <i>porV</i> (<i>fjoh_1555</i>)	This study
FI_038	FI_012 Δ <i>fjoh_1759</i> (<i>plug</i>)	This study
FI_040	FI_012 Δ <i>fjoh_4997</i> (<i>ppi</i>)	This study
FI_058	FI_012 Δ <i>porV</i> (<i>fjoh_1555</i>) Δ <i>fjoh_1759</i> (<i>plug</i>)	This study
FI_068	FI_012 Δ <i>porV</i> (<i>fjoh_1555</i>) Δ <i>fjoh_1759</i> (<i>plug</i>) Δ <i>sprA</i>	This study
Plasmid	Description ^a	Reference
pGEM-T Easy	General cloning vector; Ap ^r	Promega
pYT313	<i>sacB</i> -containing mobilizable suicide vector; Ap ^r (Em ^r)	35
pHTC HaloTag® CMV-neo	<i>Halotag</i> -containing plasmid	Promega
pET22b(+)	<i>E. coli</i> expression vector; Ap ^r	Merck
pFL53	1.8-kbp fragment of <i>chiA</i> inserted between the NdeI and XhoI sites of pET22b(+)	This study
pFL56	2.5-kbp upstream of <i>sprA</i> (including the first 63bp of <i>sprA</i>) inserted between the NcoI and SpeI sites of pGEM-T Easy	This study
pFL57	2.7-kbp of <i>sprA</i> (bp 55 to 2770) inserted between the SpeI and SacI sites of pFL56	This study
pFL58	Construct used to delete <i>sprA</i> ; contains 2.5-kbp upstream and 2.5-kbp downstream of <i>sprA</i> in pYT313	This study
pFL61	<i>halotag</i> inserted between the BamHI and SpeI sites of pFL57	This study
pFL64	<i>halotag::sprA</i> sequence from pFL61 in pYT313	This study
pFL66	<i>twin-strep-tag</i> inserted between the BamHI and SpeI sites of pFL57	This study
pFL67	<i>twin-strep-tag::sprA</i> sequence from pFL66 in pYT313	This study
pFL80	Construct used to delete <i>porV</i> (<i>fjoh_1555</i>); 2.5-kbp upstream and 2.5-kbp downstream of <i>porV</i> in pYT313	This study
pFL81	Construct used to delete <i>fjoh_1759</i> (<i>plug</i>); 2.6-kbp upstream and 2.6-kbp downstream of <i>fjoh_1759</i> in pYT313	This study
pFL82	Construct used to delete <i>fjoh_4997</i> (<i>ppi</i>); 2.6-kbp upstream and 2.7-kbp downstream of <i>fjoh_4997</i> in pYT313	This study

Some strains and plasmids have been published previously^{35,36,49}.

^aSelection markers functional in *F. johnsoniae* are in parentheses.

An experiment to search for dark-matter interactions using sodium iodide detectors

The COSINE-100 Collaboration*

Observations of galaxies and primordial radiation suggest that the Universe is made mostly of non-luminous dark matter^{1,2}. Several new types of fundamental particle have been proposed as candidates for dark matter³, such as weakly interacting massive particles (WIMPs)^{4,5}. These particles would be expected to interact with nuclei in suitable detector materials on Earth, for example, causing them to recoil. However, no definitive signal from such dark-matter interactions has been detected despite concerted efforts by many collaborations⁶. One exception is the much-debated claim by the DAMA collaboration of a statistically significant (more than nine standard deviations) annual modulation in the rate of nuclear interaction events. Annual modulation is expected because of the variation in Earth's velocity relative to the Galaxy's dark-matter halo that arises from Earth's orbital motion around the Sun. DAMA observed a modulation in the rate of interaction events in their detector^{7–9} with a period and phase consistent with that expected for WIMPs^{10–12}. Several groups have been working to develop experiments with the aim of reproducing DAMA's results using the same target medium (sodium iodide)^{13–17}. To determine whether there is evidence for an excess of events above the expected background in sodium iodide and to look for evidence of an annual modulation, the COSINE-100 experiment uses sodium iodide as the target medium to carry out a model-independent test of DAMA's claim. Here we report results from the initial operation of the COSINE-100 experiment related to the first task^{18,19}. We observe no excess of signal-like events above the expected background in the first 59.5 days of data from COSINE-100. Assuming the so-called standard dark-matter halo model, this result rules out WIMP-nucleon interactions as the cause of the annual modulation observed by the DAMA collaboration^{20–23}. The exclusion limit on the WIMP-sodium interaction cross-section is $1.14 \times 10^{-40} \text{ cm}^2$ for 10-GeV c^{-2} WIMPs at a 90% confidence level. The COSINE-100 experiment will continue to collect data for two more years, enabling a model-independent test of the annual modulation observed by the DAMA collaboration.

COSINE-100 is located at the Yangyang Underground Laboratory in South Korea and began collecting data in 2016. The experiment uses eight thallium-doped sodium iodide crystals, arranged in a 4×2 array, with a total mass of 106 kg. The crystals were grown especially for the experiment to contain low levels of radioactive contaminants. Each crystal is coupled to two photomultiplier tubes (PMTs) to measure the amount of energy deposited in the crystal by a particle interaction. The sodium iodide crystal assemblies are immersed in 2,200 l of liquid scintillator, which enables the identification and subsequent reduction of radioactive backgrounds detected by the crystals²⁴. The liquid scintillator is surrounded by copper, lead and plastic scintillator to reduce the background contribution from external radiation and cosmic-ray muons²⁵ (Extended Data Fig. 1).

The data used in this analysis were acquired between 20 October 2016 and 19 December 2016, with a total exposure of 59.5 live days. During this two-month period, no substantial environmental abnormalities or unstable detector performance were observed. The analysis

was performed with all eight crystals. Six of the crystals have light yields of about 15 photoelectrons per kiloelectronvolt, with an analysis threshold of 2 keV. The other two crystals have lower light yields and require higher analysis thresholds (4 keV and 8 keV)¹⁸. Because the direct effect of these two crystals on the experiment is not substantial, here we discuss the spectra of only the six crystals with lower thresholds. When both PMTs on the same crystal register signals that are consistent with at least one photoelectron within 200 ns, that crystal is considered to have registered a 'hit'. The outputs of all of the detector elements during 8- μ s time windows surrounding the hit time are recorded.

A nucleus recoiling from an interaction with a WIMP is expected to produce a hit in a single crystal. We select a set of candidate events by applying several criteria to reject backgrounds. We use boosted decision trees²⁶ (BDTs; a type of multivariate machine learning algorithm) to characterize the pulse shapes of the scintillation photons to discriminate PMT-induced noise events from radiation-induced events. Events that had hits in multiple crystals, the liquid scintillator or the muon detector are also rejected as multiple-hit events. Although multiple-hit events are not used for the WIMP search, they are used to develop the event selection criteria, to determine efficiencies and to model backgrounds.

Multiple-hit events recorded during the two-week calibration campaign with a ^{60}Co source provided a large sample of Compton scattering events, in which a γ -ray from the ^{60}Co source scatters from an electron in one crystal and is detected in another crystal. The BDTs

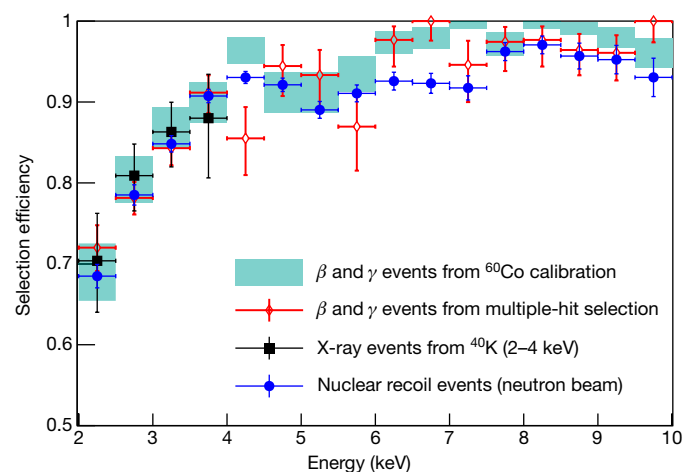


Fig. 1 | Efficiency of event selection. We use various methods to evaluate the efficiency of event selection. The statistical error bands (68% confidence interval) of the event-selection efficiencies determined from the ^{60}Co calibration data are shown as teal shaded regions and are compared with the efficiencies determined from multiple-hit events (red diamonds), internal ^{40}K coincidence events (black squares) and the nuclear-recoil calibration data (blue circles) for one of the crystals. Horizontal error bars depict the bin width of the data. Vertical error bars are 68% confidence intervals.

*A list of participants and their affiliations appears at the end of the paper.

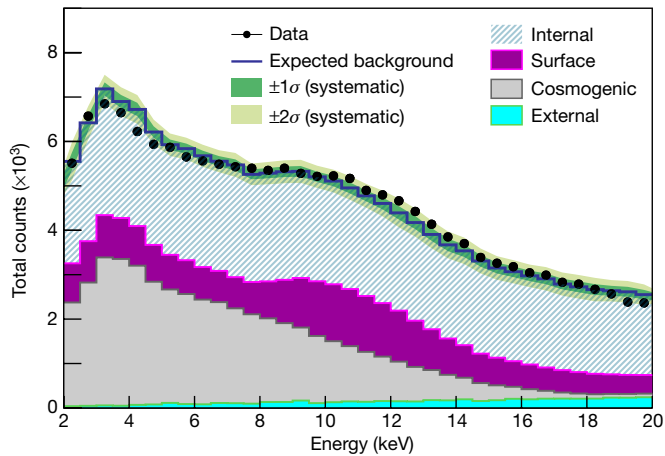


Fig. 2 | Measured and simulated energy spectra. The summed energy spectrum for the six crystals (black filled circles, with error bars (smaller than the symbol size) indicating the 68% confidence interval) and the expected background (blue line) are compared. Contributions to the background from internal radionuclide contaminations (primarily ^{210}Pb and ^{40}K), ^{210}Pb on the surfaces of the crystals and of nearby materials, cosmogenic activation (mostly ^{109}Cd and ^3H) and external contaminants (mostly ^{238}U and ^{232}Th) are indicated. The dark green (light green) band is the 68% (95%) confidence interval for the background model. The counts are shown in bins of 0.5 keV.

are trained for each detector using the multiple-hit events from the ^{60}Co calibration data—weighted to match the energy spectrum of the expected background—and physics data for the signals and the PMT-induced noise (see Methods). The efficiencies of the selection requirements are initially measured with the multiple-hit events from the ^{60}Co source (Fig. 1).

Multiple-hit events from ^{40}K decay are produced when a 3-keV X-ray registers in one crystal and its accompanying 1,460-keV γ -ray registers in another¹⁷. These events occur throughout the data exposure time and provide independent, real-time energy calibrations and efficiency measurements in the region of interest for the WIMP search (2–6 keV). The efficiencies measured with the multiple-hit events that occur during the dark-matter-search exposure, including tagged 3-keV X-rays from ^{40}K , are in agreement with the efficiencies measured using the ^{60}Co data. A specialized apparatus that has a monoenergetic 2.42-MeV neutron beam is used to measure the selection efficiencies of nuclear recoil events. This measurement was performed with a small test crystal that was cut from the same ingot as one of the crystals used for the COSINE-100 experiment. The efficiencies determined from the different methods are mutually consistent within a 5% level of uncertainty (Fig. 1). The efficiency uncertainties are included as a systematic error.

The remaining dark-matter-search data originate predominantly from environmental γ and β radiation produced from the crystals themselves or from the nearby surrounding materials. Sources include radioactive contaminants inside the crystals or on their surfaces, external detector components and cosmogenic activation¹⁹. The background spectrum for each individual crystal is modelled using simulations based on the Geant4 toolkit²⁷. Multiple-hit events with measured energies between 2 keV and 2,000 keV and single-hit events with measured energies between 6 keV and 2,000 keV are used in the modelling, as described in detail elsewhere¹⁹ (see also Methods). Single-hit events with energies below 6 keV are excluded to avoid a bias against dark-matter signal events. In Fig. 2 we show the summed single-hit event spectrum between 2 keV and 20 keV for the six crystals compared with the simulated contributions from various sources. The data in the 2–6-keV region of interest are within the error bands of the background model.

Several sources of systematic uncertainty were identified and included in the analysis. The largest uncertainties are those associated with the efficiency, which include statistical errors in the efficiency

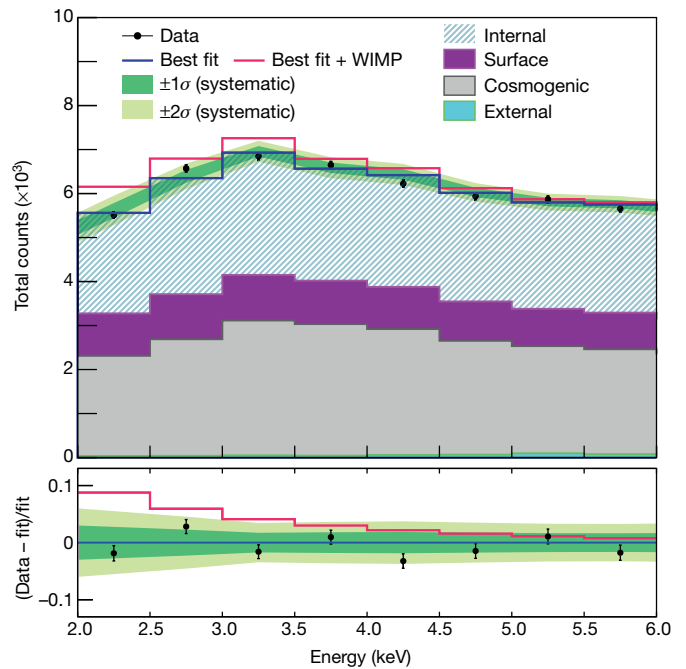


Fig. 3 | Fit results for a WIMP mass of $10 \text{ GeV } c^{-2}$. The data points (black filled circles, with error bars indicating the 68% confidence interval) show the summed energy spectra from the six crystals and the solid blue line shows the result for the fit assuming a WIMP mass of $10 \text{ GeV } c^{-2}$. The expected signal excess above the background for a WIMP mass of $10 \text{ GeV } c^{-2}$ and a spin-independent WIMP–nucleon cross-section of $2.35 \times 10^{-40} \text{ cm}^2$ is shown as a solid red line. Coloured regions are as in Fig. 2. The lower panel shows the residuals between the data and the best fit, normalized by the best fit (black filled circles). The bands of systematic uncertainty (dark and light green) and the expected DAMA/LIBRA-phase1 signal spectrum (red) are similarly shown.

determination with the ^{60}Co calibration and systematic errors derived from the independent cross-checks. Uncertainties in the energy resolution and nonlinear responses of the sodium iodide crystals²⁸ affect the shapes of the background and signal spectra. These uncertainties are studied using tagged 3-keV X-rays from internal ^{40}K and 59.5-keV γ -rays from an external ^{241}Am source. We also account for different models for ^{210}Pb decays¹⁹ and variations in the levels of external uranium and thorium decay-chain contaminants, as well as the effects of event-rate variations and possible distortions in the shapes of the background model components (Methods).

We used the simulated data to determine the contributions of dark-matter-induced nuclear recoils to the measured energy spectra. Samples of WIMP–sodium and WIMP–iodine spin-independent scattering events were generated for 18 different WIMP masses, ranging from $5 \text{ GeV } c^{-2}$ to $10,000 \text{ GeV } c^{-2}$, using the standard WIMP halo model with the same parameters that were used for the WIMP interpretation of the DAMA/LIBRA-phase1 signal¹⁰. These events were then processed through the detector simulation and the output events were subjected to the same selection criteria that were applied to the data.

To search for evidence of dark-matter-induced events, we performed binned maximum-likelihood fits to the measured single-hit energy spectra between 2 keV and 20 keV for each of the 18 WIMP masses. We used the Bayesian analysis toolkit²⁹ with probability density functions based on the shapes of the simulated WIMP signal spectra and the various components of the background model. Uniform priors were used for the signals and Gaussian priors were used for the background, with means and uncertainties for each background component set at the values determined from the model fitted to the data¹⁹. The systematic uncertainties are included in the fit as nuisance parameters with Gaussian priors. To be conservative in the assignment of systematic uncertainties, we consider the maximum allowed distortions of the shapes of the

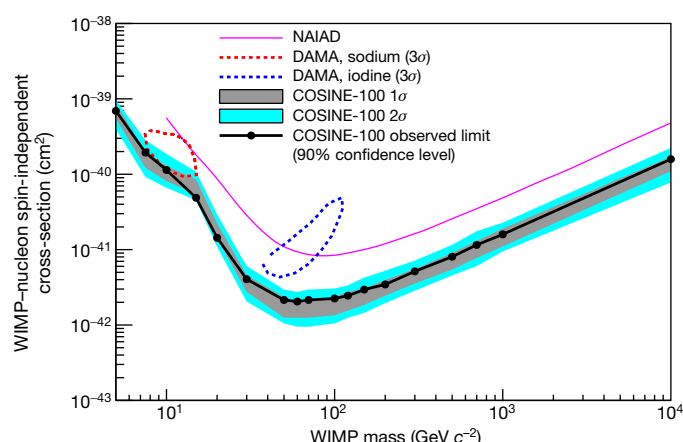


Fig. 4 | Exclusion limits on the WIMP–nucleon spin-independent cross-section. The 90% confidence level exclusion limits on the WIMP–nucleon spin-independent cross-section determined from the data from the first 59.5 days of the COSINE-100 experiment (filled circles and black solid line; total exposure of 6,303.9 kg d) are shown together with their 68% (grey shading) and 95% (blue shading) probability bands assuming the background-only hypothesis. Our exclusion limits are compared with 3σ allowed regions of the WIMP mass and the cross-section associated with the DAMA/LIBRA-phase1 signal for the WIMP–sodium (red dotted contour) and the WIMP–iodine (blue dotted contour) scattering hypothesis¹⁰. The limit from NAIAD³⁰—the only other sodium-iodide-based experiment to set a competitive limit—is shown in purple.

probability density functions within their uncertainties. We also consider the possibility of correlated rate and shape uncertainties and the uncorrelated bin-by-bin statistical uncertainties (Methods). To calculate the expected 90% confidence level upper limits on WIMP–nucleon scattering cross-sections, we performed 1,000 simulated experiments with the expected backgrounds and no dark-matter signal.

Data were fitted to each of the 18 WIMP masses. An example of a maximum-likelihood fit with a $10 \text{ GeV } c^{-2}$ WIMP signal is presented in Fig. 3 (see also Extended Data Fig. 5). The summed event spectrum for the six crystals is shown together with the best-fit result. For comparison, the expected signal for a $10 \text{ GeV } c^{-2}$ WIMP with a spin-independent cross-section of $2.35 \times 10^{-40} \text{ cm}^2$ —the central value of the DAMA/LIBRA-phase1 signal interpreted as a WIMP–sodium interaction—is overlaid in red. No excess of events that could be attributed to standard-halo WIMP interactions are found for the 18 WIMP masses considered. The posterior probabilities of the existence of a WIMP-induced signal are consistent with zero in all cases; we determined 90% confidence level limits. In Fig. 4 we show the 3σ contours of the allowed WIMP mass and the cross-sections that are associated with the DAMA/LIBRA-phase1 signal¹⁰, together with the 90% confidence level upper limits from the COSINE-100 data.

Despite strong evidence for its existence, the identity of dark matter remains a mystery. COSINE-100 continues to collect data, and several years of data will be necessary to fully confirm or refute DAMA's results. However, the first 59.5 days of background data show that the annual modulation in the signal observed by DAMA is inconsistent with spin-independent interactions between WIMPs and sodium or iodine in the context of the standard halo model.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0739-1>.

Received: 27 April 2018; Accepted: 13 September 2018;
Published online 5 December 2018.

1. Clowe, D. et al. A direct empirical proof of the existence of dark matter. *Astrophys. J.* **648**, L109–L113 (2006).

2. Ade, P. A. R. et al. Planck 2015 results. XIII. Cosmological parameters. *Astron. Astrophys.* **594**, A13 (2016).
3. Baer, H., Choi, K.-Y., Kim, J. E. & Roszkowski, L. Dark matter production in the early Universe: beyond the thermal WIMP paradigm. *Phys. Rep.* **555**, 1–60 (2015).
4. Lee, B. W. & Weinberg, S. Cosmological lower bound on heavy-neutrino masses. *Phys. Rev. Lett.* **39**, 165–168 (1977).
5. Goodman, M. W. & Witten, E. Detectability of certain dark matter candidates. *Phys. Rev. D* **31**, 3059–3063 (1985).
6. Battaglieri, M. et al. US cosmic visions: new ideas in dark matter 2017: community report. Preprint at <https://arxiv.org/abs/1707.04591> (2017).
7. Bernabei, R. et al. Searching for WIMPs by the annual modulation signature. *Phys. Lett. B* **424**, 195–201 (1998).
8. Bernabei, R. et al. Final model independent result of DAMA/LIBRA-phase1. *Eur. Phys. J. C* **73**, 2648 (2013).
9. Bernabei, R. et al. First model independent results from DAMA/LIBRA-phase2. Preprint at <https://arxiv.org/abs/1805.10486> (2018).
10. Savage, C., Gelmini, G., Gondolo, P. & Freese, K. Compatibility of DAMA/LIBRA dark matter detection with other searches. *J. Cosmol. Astropart. Phys.* **4**, 10 (2009).
11. Baum, S., Freese, K. & Kelso, C. Dark matter implications of DAMA/LIBRA-phase2 results. Preprint at <https://arxiv.org/abs/1804.01231> (2018).
12. Kang, S., Scopel, S., Tomar, G. & Yoon, J.-H. DAMA/LIBRA-phase2 in WIMP effective models. *J. Cosmol. Astropart. Phys.* **7**, 16 (2018).
13. Barbosa de Souza, E. et al. First search for a dark matter annual modulation signal with NaI(Tl) in the Southern Hemisphere by DM-Ice17. *Phys. Rev. D* **95**, 032006 (2017).
14. Amaré, J. et al. Status of the ANAIS dark matter project at the Canfranc Underground Laboratory. *J. Phys. Conf. Ser.* **718**, 042052 (2016).
15. Fushimi, K. et al. Dark matter search project PICO-LON. *J. Phys. Conf. Ser.* **718**, 042022 (2016).
16. Xu, J., Calaprice, F., Froberg, F., Shields, E. & Suerfer, B. SABRE – a test of DAMA with high-purity NaI(Tl) crystals. *AIP Conf. Proc.* **1672**, 040001 (2015).
17. Adhikari, P. et al. Understanding internal backgrounds in NaI(Tl) crystals toward a 200 kg array for the KIMS-Nal experiment. *Eur. Phys. J. C* **76**, 185 (2016).
18. Adhikari, G. et al. Initial performance of the COSINE-100 experiment. *Eur. Phys. J. C* **78**, 107 (2018).
19. Adhikari, P. et al. Background model for the NaI(Tl) crystals in COSINE-100. *Eur. Phys. J. C* **78**, 490 (2018).
20. Tanabashi, M. et al. The review of particle physics. *Phys. Rev. D* **98**, 030001 (2018).
21. Drukier, A. K., Freese, K. & Spergel, D. N. Detecting cold dark-matter candidates. *Phys. Rev. D* **33**, 3495–3508 (1986).
22. Freese, K., Frieman, J. A. & Gould, A. Signal modulation in cold dark matter detection. *Phys. Rev. D* **37**, 3388–3405 (1988).
23. Lewin, J. & Smith, P. Review of mathematics, numerical factors, and corrections for dark matter experiments based on elastic nuclear recoil. *Astropart. Phys.* **6**, 87–112 (1996).
24. Park, J. S. et al. Performance of a prototype active veto system using liquid scintillator for a dark matter search experiment. *Nucl. Instrum. Methods A* **851**, 103–107 (2017).
25. Prihadi, H. et al. Muon detector for the COSINE-100 experiment. *J. Instrum.* **13**, T02007 (2018).
26. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
27. Agostinelli, S. et al. GEANT4: a simulation toolkit. *Nucl. Instrum. Methods A* **506**, 250–303 (2003).
28. Swiderski, L. Response of doped alkali iodides measured with gamma-ray absorption and Compton electrons. *Nucl. Instrum. Methods A* **705**, 42–46 (2013).
29. Caldwell, A., Kollár, D. & Kröninger, K. BAT – the Bayesian analysis toolkit. *Comput. Phys. Commun.* **180**, 2197–2209 (2009).
30. Alner, G. J. et al. Limits on WIMP cross-sections from the NAIAD experiment at the Boulby Underground Laboratory. *Phys. Lett. B* **616**, 17–24 (2005).

Acknowledgements We thank the Korea Hydro and Nuclear Power (KHNP) Company for providing underground laboratory space at Yangyang. This work is supported by: the Institute for Basic Science (IBS) under project code IBS-R016-A1 and NRF-2016R1A2B3008343, South Korea; UIUC campus research board, the Alfred P. Sloan Foundation Fellowship, NSF grant numbers PHY-1151795, PHY-1457995, DGE-1122492 and DGE-1256259, WIPAC, the Wisconsin Alumni Research Foundation, Yale University and DOE/NNSA grant number DE-FC52-08NA28752, USA; STFC grants ST/N000277/1 and ST/K001337/1, UK; and CNPq and grant number 2017/02952-0 FAPESP, Brazil.

Reviewer information Nature thanks B. Sadoulet and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Y.K., H.S.L., R.H.M. and N.J.C.S. conceived the COSINE-100 experiment. Its design and installation were led by K.P. and C.H.H. and carried out by all members of the collaboration. Operation and maintenance were organized by C.H.H. with support from on-site crews, W.G.K., B.K. and S.H.Y. Jaeson Lee, J.P., J.H.J., G.A., P.A., H. Prihadi, C.H.H., W.G.T., E.B.d.S., H.S.L. and K.K. contributed to data acquisition, production and verification. H.J., Hyeonseo Park and K.K. provided nuclear recoil data. P.A., G.A., J.P., K.K., H. Prihadi, N.Y.K. and C.H.H. performed the source calibrations. Hyounggyu Kim, N.Y.K., C.H.H. and H.S.L. developed the slow

control framework. J.H.J. and W.G.T. developed the data monitoring package. N.Y.K., Jooyoung Lee and Y.J.K. provided the radiopurity of the detector materials. G.A., J.P. and N.Y.K. produced the liquid scintillator. Background simulations were performed by F.M., E.J., P.A., W.G.T. and E.B.d.S. C.H.H. and P.A. analysed the observational and simulated data. The manuscript and plots were produced by C.H.H. and H.S.L., and edited by R.H.M., S.L.O., N.J.C.S. and the other members of the collaboration. All authors participated in online data-monitoring shifts and approved the manuscript. Authors are listed alphabetically by their last names.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0739-1>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0739-1>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to C.H.H. and H.S.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

The COSINE-100 Collaboration

Govinda Adhikari¹, Pushparaj Adhikari¹, Estella Barbosa de Souza², Nelson Carlin³, Seonho Choi⁴, Mitra Djamal⁵, Anthony C. Ezeribe⁶, Chang Hyon Ha^{7*}, Insik Hahn⁸, Antonia J. F. Hubbard^{2,16}, Eunju Jeon⁷, Jay Hyun Jo², Hanwool Joo⁴, Woon Gu Kang⁷, Woosik Kang⁹, Matthew Kauer¹⁰, Bonghee Kim⁷, Hongjoo Kim¹¹, Hyounggyu Kim⁷, Kyungwon Kim⁷,

Nam Young Kim⁷, Sun Kee Kim⁴, Yeongduk Kim^{1,7}, Yong-Hamb Kim^{7,12}, Young Ju Ko⁷, Vitaly A. Kudryavtsev⁶, Hyun Su Lee^{7*}, Jaison Lee⁷, Jooyoung Lee¹¹, Moo Hyun Lee⁷, Douglas S. Leonard⁷, Warren A. Lynch⁶, Reina H. Maruyama², Frederic Mouton⁶, Stephen L. Olsen⁷, Byungju Park¹³, Hyang Kyu Park¹⁴, Hyeonseo Park¹², Jungsic Park^{7,17}, Kangsoon Park⁷, Walter C. Pettus^{2,18}, Hafizh Prihadi⁵, Sejin Ra⁷, Carsten Rott⁹, Andrew Scarff^{6,19}, Keon Ah Shin⁷, Neil J. C. Spooner⁶, William G. Thompson², Liang Yang¹⁵ & Seok Hyun Yong⁷

¹Department of Physics, Sejong University, Seoul, South Korea. ²Department of Physics, Yale University, New Haven, CT, USA. ³Physics Institute, University of São Paulo, São Paulo, Brazil. ⁴Department of Physics and Astronomy, Seoul National University, Seoul, South Korea. ⁵Department of Physics, Bandung Institute of Technology, Bandung, Indonesia. ⁶Department of Physics and Astronomy, University of Sheffield, Sheffield, UK. ⁷Center for Underground Physics, Institute for Basic Science (IBS), Daejeon, South Korea. ⁸Department of Science Education, Ewha Womans University, Seoul, South Korea. ⁹Department of Physics, Sungkyunkwan University, Suwon, South Korea. ¹⁰Department of Physics and Wisconsin IceCube Particle Astrophysics Center, University of Wisconsin-Madison, Madison, WI, USA. ¹¹Department of Physics, Kyungpook National University, Daegu, South Korea. ¹²Korea Research Institute of Standards and Science, Daejeon, South Korea. ¹³IBS School, University of Science and Technology (UST), Daejeon, South Korea. ¹⁴Department of Accelerator Science, Korea University, Sejong, South Korea. ¹⁵Department of Physics, University of Illinois at Urbana-Champaign, Urbana, IL, USA. ¹⁶Present address: Department of Physics and Astronomy, Northwestern University, Evanston, IL, USA. ¹⁷Present address: High Energy Accelerator Research Organization (KEK), Tsukuba, Japan. ¹⁸Present address: Center for Experimental Nuclear Physics and Astrophysics, Department of Physics, University of Washington, Seattle, WA, USA. ¹⁹Present address: Department of Physics and Astronomy, University of British Columbia, Vancouver, British Columbia, Canada. *e-mail: changhyon.ha@gmail.com; hyunsulee@ibs.re.kr

METHODS

The COSINE-100 experiment is located 700 m below the ground at the Yangyang Underground Laboratory in eastern South Korea. A cut-out view of the detector is shown in Extended Data Fig. 1. It consists of an array of eight sodium iodide NaI(Tl) scintillating crystals (total mass of 106 kg) immersed in 2,200 l of liquid scintillator contained in an acrylic box that is surrounded by copper and lead shielding. Plastic scintillators surround the entire apparatus to detect cosmic-ray muons that penetrate the apparatus. External radiation is attenuated by the lead, copper and liquid scintillator shields. Signals from the liquid scintillator and muon detectors are used to identify background events that are induced by radiation sources in or near the crystals and by cosmic-ray muons. More details about the experimental site, including the fluxes of cosmic-ray muons and neutrons, and the data acquisition system can be found elsewhere^{18,25,31}.

Event selection. Pulse shapes from the detector are recorded when both PMTs on a crystal record signals that correspond to at least one single photoelectron within 200 ns. In the offline analysis, events are rejected if they occur within 30 ms of a signal from any of the surrounding muon detectors or if there is a signal in the liquid scintillator within 4 μs . Events with and without accompanying hit crystals in an 8- μs time window are classified as multiple-hit and single-hit events, respectively. Events are further classified according to their energy: 2–70 keV is low energy and 70–2,000 keV is high energy.

Extended Data Fig. 2a shows an averaged waveform for radiation-induced scintillation light signals in the NaI(Tl) crystal detectors, where the characteristic 250-ns NaI(Tl) scintillation light decay time is evident. By contrast, PMT noise pulses, which are considerably more frequent, decay faster, with decay times ranging between 20 ns and 50 ns (Extended Data Fig. 2b). Some detectors intermittently produce events that have slow rise and decay times (Extended Data Fig. 2c); these are attributed to PMT discharges.

We use BDTs to separate signal events from noise events. The fast PMT noise-induced events are efficiently removed by a BDT that is based on the amplitude-weighted average time of a signal, the ratios of the leading-edge and trailing-edge charge sums relative to total charge, and the balance of deposited energies between the two PMTs. This BDT is trained with a sample of signal-rich, energy-weighted, multiple-hit events from the ^{60}Co calibration and single-hit events from the WIMP-search data; the latter are mostly triggered by PMT noise. A second BDT (BDTA) that includes weighted higher-order time moments is effective at eliminating discharge events. Extended Data Fig. 3 shows two-dimensional scatter plots of the BDT and BDTA outputs for two separate crystals, one with and the other without PMT discharge signals. Events that are above and to the right of the dashed red lines in the figure are retained.

Background modelling. The primary background components of the energy spectra of the crystals are from internal ^{238}U , ^{232}Th , ^{40}K and ^{210}Pb contaminations in the bulk material of the crystal, plus additional ^{210}Pb on the surfaces of the crystal and its reflective wrapping foil, caused by exposure to atmospheric radon during the encapsulation of the crystals¹⁹. In addition, we considered background from external sources such as ^{238}U , ^{232}Th and ^{40}K contaminations in the PMTs, the liquid scintillator and the bulk material of the surrounding shields. The modelling of these contributions used starting values based on radioassay results from an underground, high-purity Ge detector³². The modelling of contributions from cosmogenic activity in the crystals was guided by measured surface production rates in NaI(Tl) ³³ and the above- and below-ground histories of each individual crystal.

In 10.7% of ^{40}K decays, a roughly 3-keV K-shell X-ray (or Auger electron) is produced in coincidence with a 1,460-keV γ -ray. Because this results in a peaking background in the WIMP search region of interest, it is of particular concern. However, in the COSINE-100 detector, about 80% of these 3-keV X-rays are tagged by the detection of its accompanying 1,460-keV γ -ray in one of the other crystals or in the liquid scintillator and so can be vetoed. The measured rate for these tagged events is used to establish the contribution of untaged 3-keV ^{40}K -induced events to the background in the region of interest for the single-hit spectrum in each crystal.

Extended Data Fig. 4 shows the results of the model fits to the data for the four categories of events (single-hit and multiple-hit events in low and high energy), with 1σ and 2σ uncertainty bands indicated in green and yellow, respectively¹⁹. All four distributions were fitted simultaneously. To avoid biasing the WIMP search, the 2–6-keV region of the low-energy, single-hit spectrum (Extended Data Fig. 4a) is not included in the fitting. The fitted model indicates that the main contributions in the 2–6-keV region of interest are from internal ^{40}K and ^{210}Pb , and cosmogenic ^{109}Cd and ^3H . The ^{109}Cd contribution was confirmed independently by a time-dependent analysis.

Systematic uncertainties. The results of the analysis are limited by the systematic uncertainties. Errors in the selection efficiency, energy resolution, energy scale and background modelling technique translate into uncertainties in the shapes of the probability density functions of the signal and background components that are used in the likelihood fit, and thus affect the results. These quantities are allowed to vary within their errors in the likelihood as nuisance parameters. Of these, the systematic errors associated with the efficiencies have the largest effects on the results. Uncertainties in the efficiencies are determined by the statistical errors from the data from the efficiency measurements for the multiple-hit ^{60}Co source, and their stability is verified by independent datasets (Fig. 1). The efficiency systematic that maximally covers the statistical errors in the region of interest mimics the shape of a WIMP signal.

For most of the energy range, the resolutions and scales are well measured with internal radioactive peaks and external calibrations. However, because external source measurements are impractical for energies below 10 keV, the resolution and scale values for these energies are determined with the samples of tagged 3-keV X-rays from the internal ^{40}K contamination. For these, statistical errors dominate and are taken as the systematic spread from these quantities. We used changes that occur in the background model when the simulation is done with different locations of the U/Th contamination in the PMTs, and alternative Geant4 methods for X-ray production of ^{210}Pb , as the systematic error from this source. The inclusion of the total systematic uncertainties degrades the sensitivity by a factor of 2.3. **WIMP extraction Bayesian fit.** A Bayesian analysis with a likelihood formulated in equation (1) was performed and this fitter, which is more computationally demanding than the background modelling fits, was run with the WIMP-search data (low-energy single-hit spectrum) between 2 keV and 20 keV. The function that is maximized has the form

$$\mathcal{L} = \prod_i^{N_{\text{ch}}} \prod_j^{N_{\text{bin}}} \frac{\mu_{ij}^{n_{ij}}}{n_{ij}!} \exp(-\mu_{ij}) \prod_k^{N_{\text{bkg}}} \exp\left[-\frac{(x_k - \alpha_k)^2}{2\sigma_k^2}\right] \prod_l^{N_{\text{syst}}} \exp\left[-\frac{x_l^2}{2\sigma_l^2}\right] \quad (1)$$

where N_{ch} is the number of crystals, N_{bin} is the number of bins in each histogram, N_{bkg} is the number of background components, N_{syst} is the number of systematic nuisance parameters, n_{ij} is the number of observed counts and μ_{ij} is the total model expectation by summing all N_{bkg} background components and a WIMP signal component after apply a shape change due to N_{syst} systematic effects. In the first product of Gaussians, x_k is the value of the k th background component, α_k is the mean value and σ_k is its 68% error. The second product of Gaussians x_l is the l th systematic parameter and σ_l is its error.

To avoid biasing the WIMP search, the fitter was developed and tested with simulated event samples. All eight crystals are fitted simultaneously with a common WIMP-signal model for each assumed WIMP mass and with fits performed for 18 different WIMP masses between 5 $\text{GeV } c^{-2}$ and 10,000 $\text{GeV } c^{-2}$. The shapes of the energy spectra of the WIMP signal are determined from simulations based on the standard WIMP halo model with parameters taken from ref. ¹⁰. To relate the simulated WIMP signals, which are caused by nuclear recoils, to our energy scale, which is calibrated with electron recoils, we use the same NaI(Tl) quenching factors that were used in interpretations of the DAMA/LIBRA-phase1 signal, $Q_{\text{Na}} = 0.3$ and $Q_1 = 0.09$ (quenching factors are the ratio of scintillation-light energy determinations for nuclear and electron recoils of the same energy).

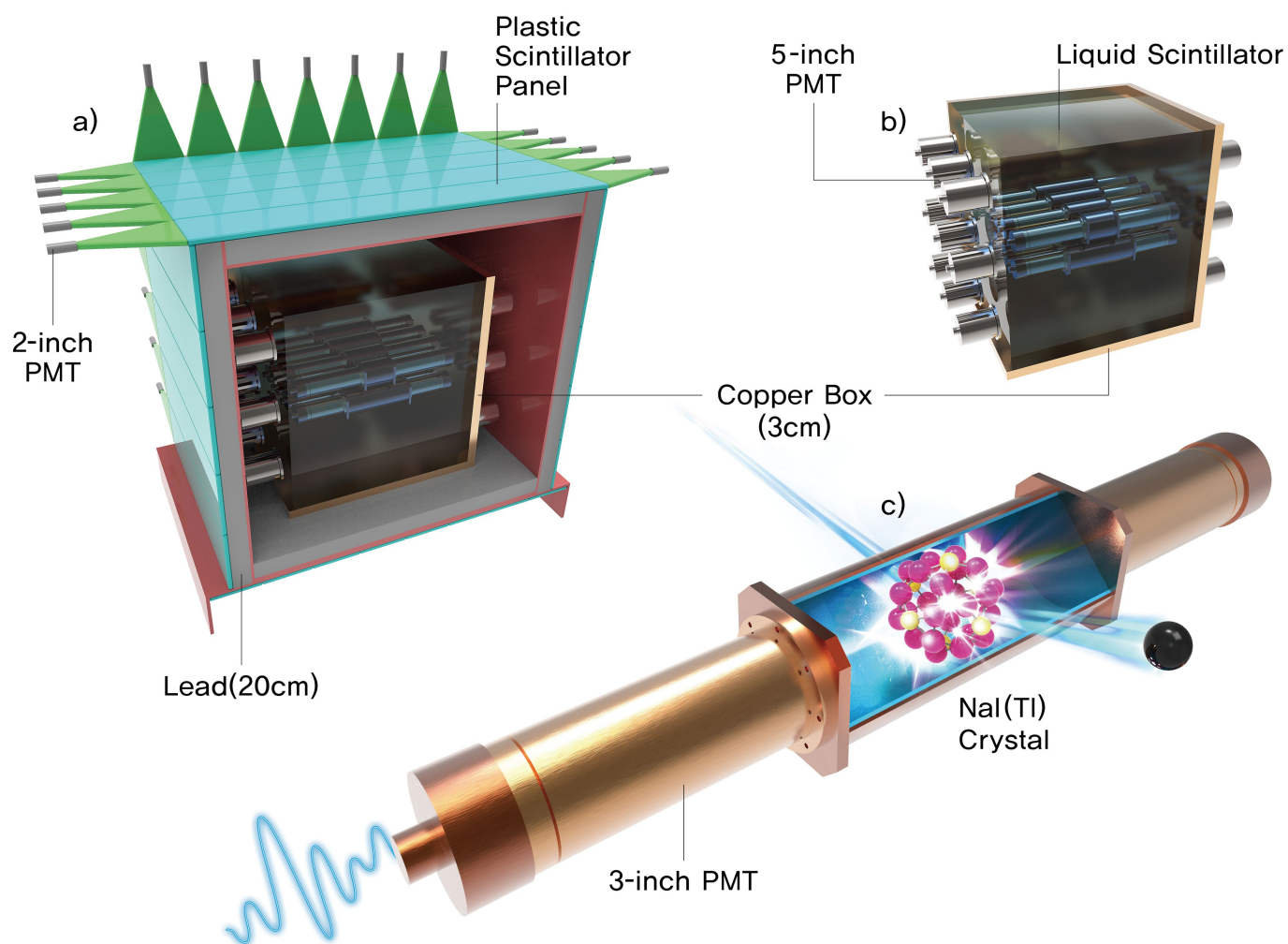
We use Gaussian priors for the normalizations of the background components and the systematic nuisance parameters for efficiencies, energy resolutions and energy scales. The initial values for the background-component normalizations are taken from the fits described above that do not use the single-hit events in the 2–6-keV region of interest. The final fit values for all nuisance parameters are within $\pm 1\sigma$ of their initial values.

Code availability. All data related to the analysis are in the ROOT (<https://root.cern.ch>) format. Analysis toolkits such as ROOT, including BDT and BAT (<https://bat.mpp.mpg.de>), are available online. Our custom codes are available from the corresponding authors on reasonable request.

Data availability

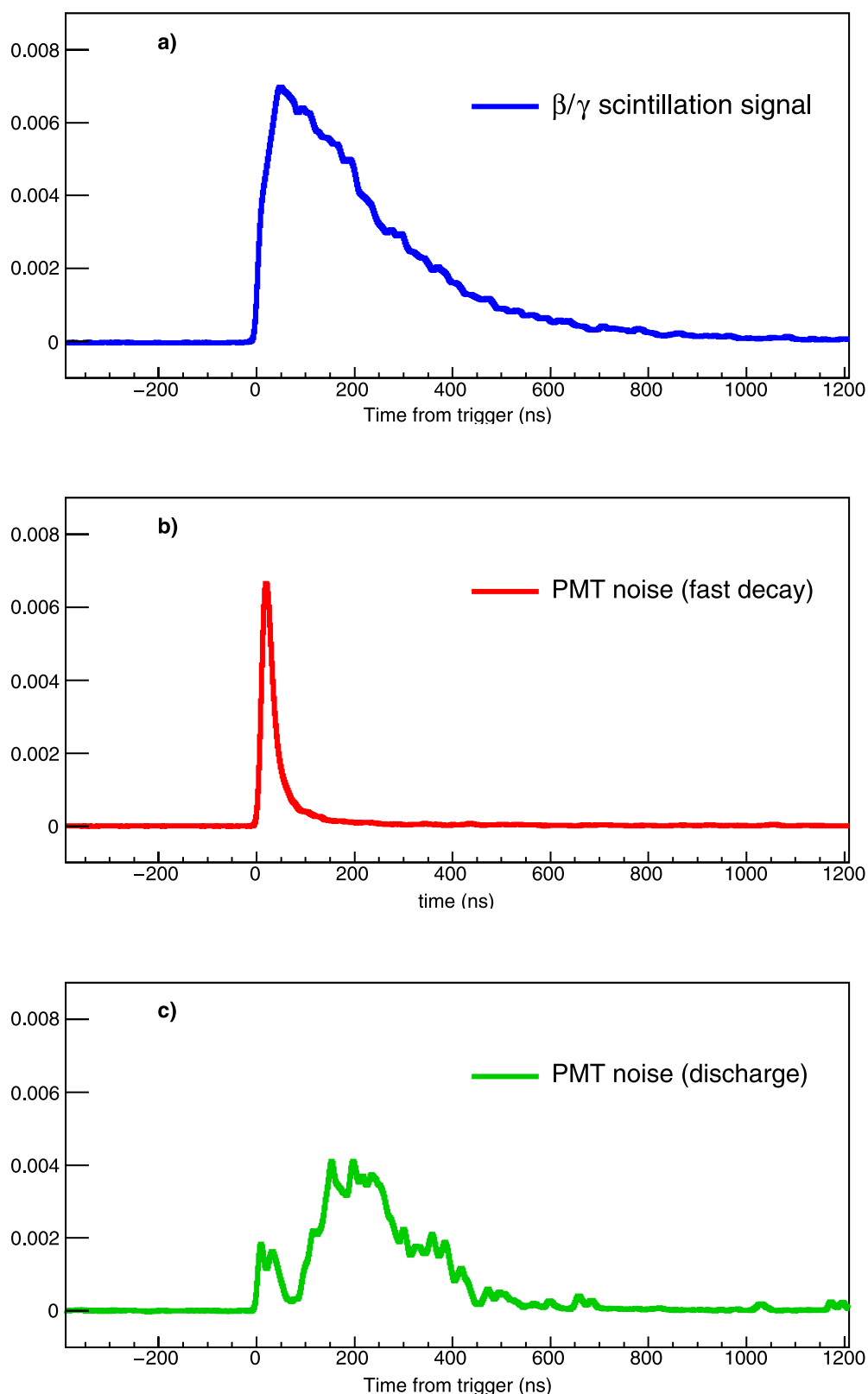
The data that support the findings of this study are available from the corresponding authors on reasonable request. Source Data for Figs. 1–4 are provided with the online version of the paper.

- Adhikari, G. et al. The COSINE-100 data acquisition system. *J. Instrum.* **13**, P09006 (2018).
- Sala, E. et al. Development of an underground low background instrument for high sensitivity measurements. *J. Phys. Conf. Ser.* **718**, 062050 (2016).
- Pettus, W. C. *Cosmogenic Activation in NaI Detectors for Dark Matter Searches*. PhD thesis, Univ. Wisconsin–Madison (2015).



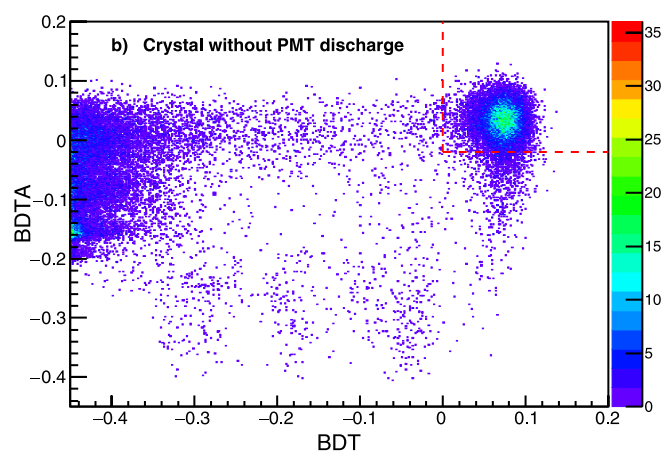
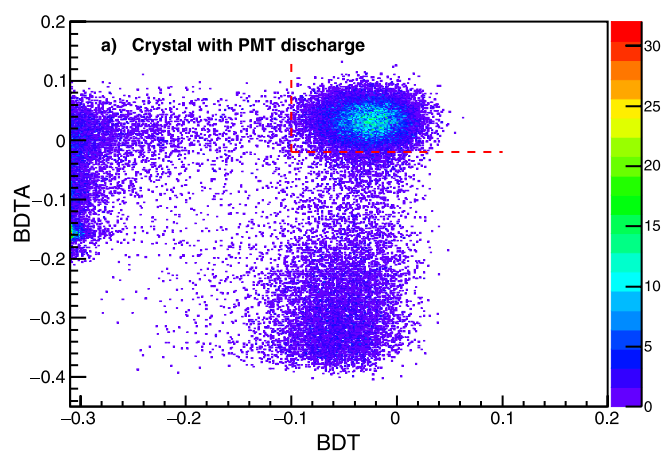
Extended Data Fig. 1 | The COSINE-100 detector. **a**, The detector is contained within a nested arrangement of shielding components, as indicated by different colours. The main purpose of the shield is to provide 4π coverage against external radiation from various background sources. The shielding components include plastic scintillator panels (blue), a lead

brick enclosure (grey) and a copper box (reddish brown). **b**, **c**, The eight encapsulated sodium iodide crystal assemblies (**c**) are located inside the copper box and are immersed in scintillating liquid (**b**). All images are schematic.



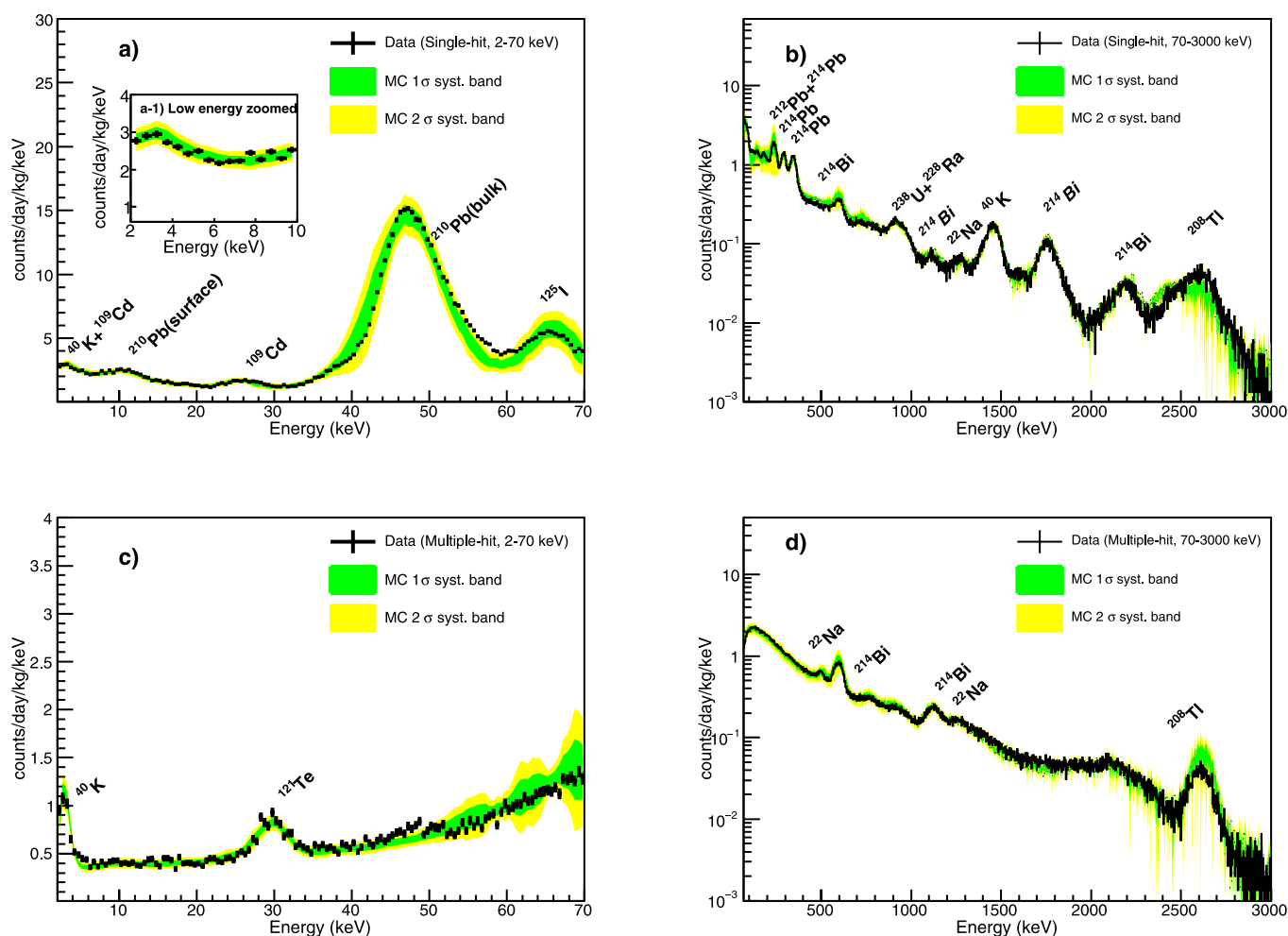
Extended Data Fig. 2 | Typical waveforms from the COSINE-100 PMTs for 2–6-keV signals. a, The β and γ scintillation signals have a fast rise and then fall off with a decay time of about 250 ns. The waveform from

WIMPs is expected to closely resemble the β and γ waveforms. **b, c,** Background waveforms from PMT noise (**b**) and external discharge (**c**).



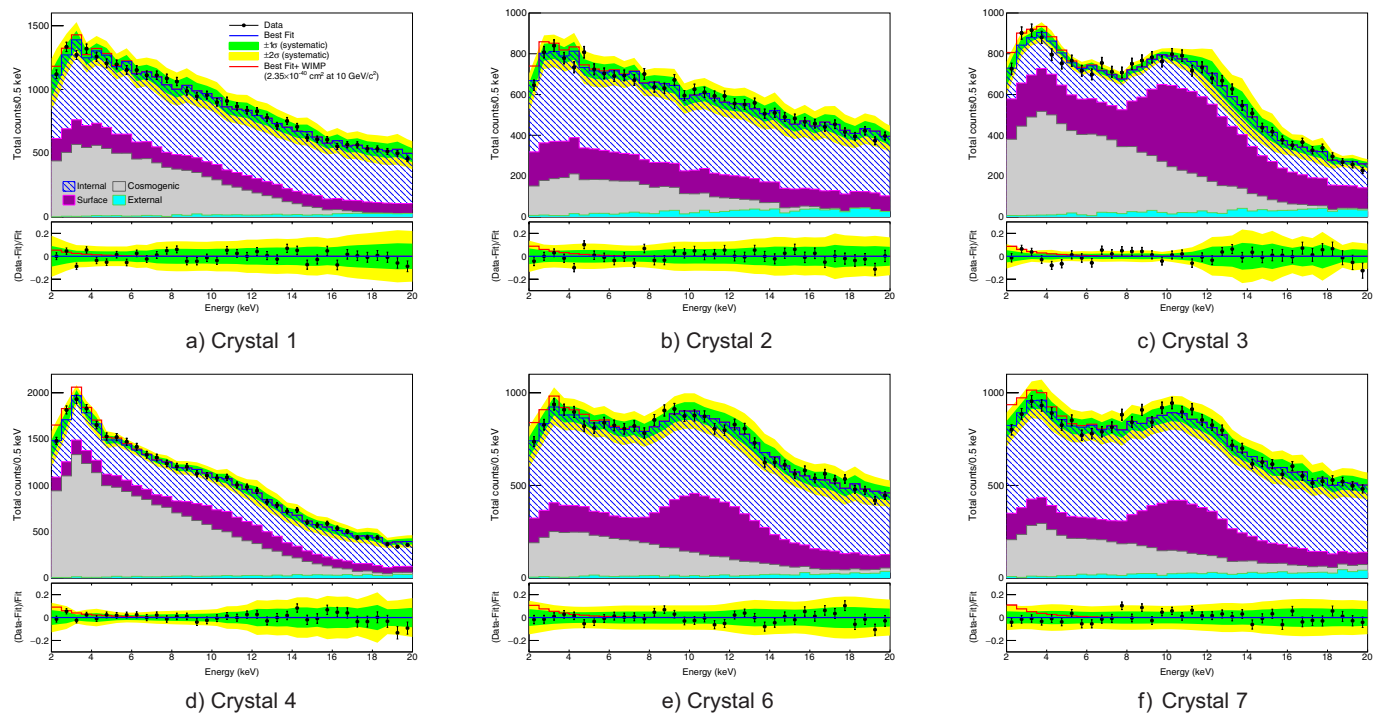
Extended Data Fig. 3 | The BDT output (horizontal) versus the BDTA output (vertical). a, b, Events (colour scale) with energies below 10 keV are shown for two separate crystals, with (a) and without (b) PMT discharge. The events to the right and above the red dotted lines are

scintillation events induced by real particle–crystal interactions. PMT noise events are to the left of the vertical dotted lines in both panels; PMT discharge events are below the horizontal dotted line in a.



Extended Data Fig. 4 | A comparison between data and simulation. **a–d**, Four categories of data are shown: single-hit low-energy (2–70 keV; **a**); single-hit high-energy (70–3,000 keV; **b**); multiple-hit low-energy (2–70 keV; **c**); and multiple-hit high-energy (70–3,000 keV; **d**). The black points (with errors bars indicating the 68% confidence interval) are data.

The green (yellow) band shows the $\pm 1\sigma$ ($\pm 2\sigma$) uncertainty range of the model. The peak near 3 keV in the multiple-hit, low-energy spectrum (**c**) is due to the tagged ^{40}K events. The inset in **a** shows a zoomed-in view in the region of interest after efficiency corrections are applied. The major contributors to the radioactive background are labelled.



Extended Data Fig. 5 | Crystal-by-crystal fit results. a–f, The points (with errors bars indicating the 68% confidence interval) show the measured energy spectra for each of the six crystals. The fit results are shown as blue histograms, with the $\pm 1\sigma$ ($\pm 2\sigma$) error bands shown in green (yellow). To compare the signal strength of the DAMA sodium

region with our data, a $10 \text{ GeV } c^{-2}$ WIMP signal at $2.35 \times 10^{-40} \text{ cm}^2$ (the centre of the DAMA sodium region) is indicated for each crystal as a red histogram. The fit residuals, together with the expectations for the $10 \text{ GeV } c^{-2}$ WIMP signal are also shown (bottom panels).

Atomic clock performance enabling geodesy below the centimetre level

W. F. McGrew^{1,2}, X. Zhang^{1,3}, R. J. Fasano^{1,2}, S. A. Schäffer^{1,4}, K. Beloy¹, D. Nicolodi^{1,2}, R. C. Brown^{1,8}, N. Hinkley^{1,2,9}, G. Milani^{1,5,6}, M. Schioppo^{1,10}, T. H. Yoon^{1,7} & A. D. Ludlow^{1,2,*}

The passage of time is tracked by counting oscillations of a frequency reference, such as Earth's revolutions or swings of a pendulum. By referencing atomic transitions, frequency (and thus time) can be measured more precisely than any other physical quantity, with the current generation of optical atomic clocks reporting fractional performance below the 10^{-17} level^{1–5}. However, the theory of relativity prescribes that the passage of time is not absolute, but is affected by an observer's reference frame. Consequently, clock measurements exhibit sensitivity to relative velocity, acceleration and gravity potential. Here we demonstrate local optical clock measurements that surpass the current ability to account for the gravitational distortion of space-time across the surface of Earth. In two independent ytterbium optical lattice clocks, we demonstrate unprecedented values of three fundamental benchmarks of clock performance. In units of the clock frequency, we report systematic uncertainty of 1.4×10^{-18} , measurement instability of 3.2×10^{-19} and reproducibility characterized by ten blinded frequency comparisons, yielding a frequency difference of $[-7 \pm (5)_{\text{stat}} \pm (8)_{\text{sys}}] \times 10^{-19}$, where 'stat' and 'sys' indicate statistical and systematic uncertainty, respectively. Although sensitivity to differences in gravity potential could degrade the performance of the clocks as terrestrial standards of time, this same sensitivity can be used as a very sensitive probe of geopotential^{5–9}. Near the surface of Earth, clock comparisons at the 1×10^{-18} level provide a resolution of one centimetre along the direction of gravity, so the performance of these clocks should enable geodesy beyond the state-of-the-art level. These optical clocks could further be used to explore geophysical phenomena¹⁰, detect gravitational waves¹¹, test general relativity¹² and search for dark matter^{13–17}.

Einstein first predicted in his general theory of relativity that gravity alters time, an effect sometimes called the gravitational redshift. Relative to a given observer, time (and the devices that measure time—clocks) is seen to evolve more slowly deeper in a gravity potential. To make meaningful comparisons between atomic clocks, this shift must be accounted for by transforming to a common reference surface, such as the geoid, which is the equipotential surface that best fits global-mean sea level of the rotating Earth. In practice, the internationally recognized coordinate time system Terrestrial Time (TT) implicitly defines a surface of constant geopotential that is near the geoid but does not change as a result of, for example, eustatic sea-level rise¹⁸. Relative heights between two 'nearby' locations (separated by as much as a few hundred kilometres) can be determined with millimetre resolution by spirit levelling¹⁹, but absolute geopotential determination can best be performed by using the Global Navigation Satellite System (GNSS) to measure ellipsoidal height and accounting for gravity by using a geoid model, leading to a total height uncertainty of several centimetres^{20,21}. Near the surface of Earth, relativistic effects amount to a fractional frequency shift of 1.1×10^{-18} per centimetre of vertical displacement.

The geopotential determination afforded by modern GNSS and gravity measurements is sufficient for state-of-the-art microwave clocks with a systematic uncertainty approaching 1×10^{-16} , corresponding to 0.9 m of elevation change, but the next generation of clocks has the potential to push the boundaries of geodetic precision. These clocks are based on optical transitions, in which greater oscillation frequency leads to an increase of 10^5 in the quality factor of the transition and concomitant performance improvements. Clock performance is characterized by systematic uncertainty (the potential error of the measured transition frequency from its unperturbed value), instability (the statistical precision that the clock affords a measurement) and reproducibility (the measured agreement between similar but distinct clocks). We report the realization of unprecedented levels in each of these three benchmarks and demonstrate in a local comparison that these clocks can now perform beyond the present-day ability to account for gravitational effects of time across the surface of Earth. This opens up the possibility of using clocks as precise, next-generation geodetic tools.

We first characterize all known sources of systematic uncertainty in two ytterbium optical lattice clocks, denoted Yb-1 and Yb-2 throughout (Fig. 1). We report a total uncertainty of 1.4×10^{-18} for each system, as shown in Table 1. Both systems exploit in-vacuum, room-temperature thermal shields surrounding the lattice-trapped atoms (pictured in Fig. 1)²². This facilitates characterization of a key systematic effect afflicting optical clocks, namely Stark effects from blackbody radiation (BBR) bathing the ultracold atoms, and further provides an in situ Faraday enclosure shielding the atoms from static electric fields due to stray charges on the vacuum apparatus²³. These clocks also utilize a one-dimensional optical lattice operating near the 'magic wavelength', at which lowest-order trap light shifts on the clock transition are cancelled, and higher-order lattice effects have been experimentally characterized²⁴. Although alternative architectures utilize cryogenic operation²⁵ or three-dimensional optical lattices^{26,27}, the unprecedented clock performance reported here requires only a one-dimensional lattice and room-temperature operation, which are useful traits for future portable apparatus or robust primary standards of time and frequency. While we treat most systematic effects in Methods, here we highlight two important effects that have not been experimentally characterized previously in optical clocks.

First, collisions of the ultracold atoms with background gases lead to a shift of the atomic transition frequency. This shift is expected to scale with the trap loss rate, Γ , although this scaling has not been experimentally confirmed previously. Furthermore, theoretical determination of the scaling coefficient requires precise knowledge of the van der Waals coefficients of the two colliding particles²⁸. The residual gas pressure in our two vacuum systems is about 0.1 μPa , with H_2 by far the dominant gas species, determined by residual gas analysis. To measure the effects of background gas, we heat a non-evaporable getter pump and induce outgassing, decreasing the trap lifetime of Yb-1 to

¹National Institute of Standards and Technology, Boulder, CO, USA. ²Department of Physics, University of Colorado, Boulder, CO, USA. ³State Key Laboratory of Advanced Optical Communication Systems and Networks, Institute of Quantum Electronics, School of Electronics Engineering and Computer Science, Peking University, Beijing, China. ⁴Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark. ⁵Istituto Nazionale di Ricerca Metrologica, Torino, Italy. ⁶Politecnico di Torino, Torino, Italy. ⁷Department of Physics, Korea University, Seoul, South Korea. ⁸Present address: Georgia Tech Research Institute, Atlanta, GA, USA. ⁹Present address: Stable Laser Systems, Boulder, CO, USA. ¹⁰Present address: National Physical Laboratory (NPL), Teddington, UK. *e-mail: andrew.ludlow@nist.gov

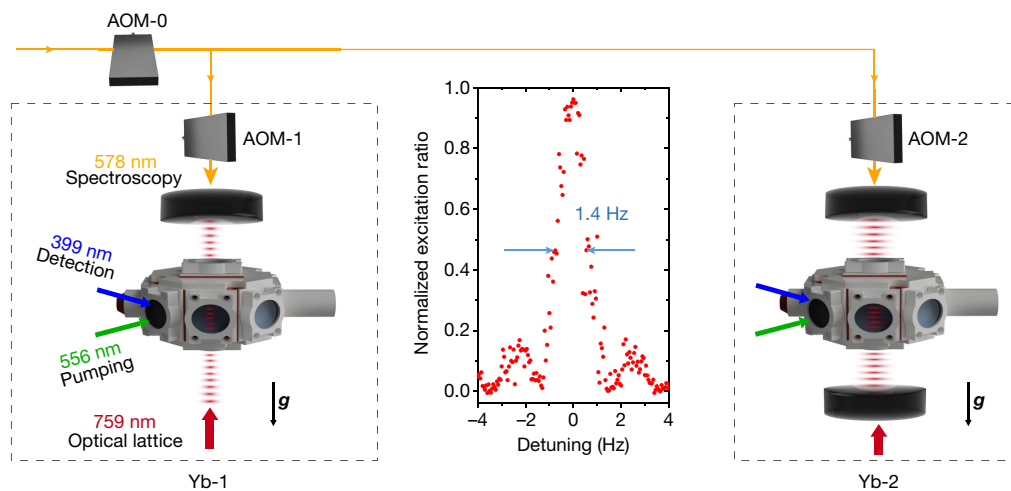


Fig. 1 | Simplified experimental scheme. ^{171}Yb atoms are cooled and loaded into two vertically oriented one-dimensional optical lattices (shown red within optical clocks Yb-1, left panel, and Yb-2, right panel). Frequency corrections to the clock laser (yellow arrows) are applied by three acousto-optic modulators (AOMs). AOM-1 and AOM-2 also cancel the optical path length fluctuations for Yb-1 and Yb-2, with the lattice mirror serving as a phase reference. The pumping laser (green arrows)

creates a spin-polarized atomic sample, and the detection laser (blue arrows) reads out atomic populations. The atoms are surrounded by an in-vacuum room-temperature thermal shield (see Methods). Middle panel, typical 560-ms Rabi spectrum with a Fourier-limited full-width at half-maximum of 1.4 Hz (arrowed). Shown is a single trace with no averaging and a signal-to-noise ratio characteristic of the measurements reported here.

as low as 93 ms. Residual gas analysis confirms that the released gas is more than 95% H_2 . We determine the trap lifetime to better than 5% uncertainty from the ratio of the atomic populations measured interleaving between two experimental cycles with variable time delay before detection. We determine the induced shift by comparing the frequency of Yb-1 with Yb-2, the latter serving as a stable frequency reference. As shown in Fig. 2a, the fractional frequency shift from background gas collisions is confirmed to vary linearly with loss rate, with a coefficient of $-1.64(12) \times 10^{-17}$ s, where the parenthetical value represents the 1σ uncertainty of the coefficient. Characteristic values of the time constant for Yb-1 and Yb-2 are 3.0 s and 4.5 s, respectively, and this value is measured every few days. The shift for Yb-1 is therefore $-5.5(5) \times 10^{-18}$, and for Yb-2, $-3.6(3) \times 10^{-18}$. Uncertainty in the

coefficient of the background gas shift is common-mode between the systems, and the differential uncertainty amounts to 3×10^{-19} , including both coefficient uncertainty and measurement uncertainty of the loss rate. We find that accurate determination of this shift is crucial for enabling 10^{-18} uncertainty with typical vacuum levels found in optical lattice clock systems. In addition, repeating this measurement at several lattice depths, up to eight times its operational value (see Methods), we observe that this shift is independent of trap depth at the 10% level.

Second, collisions between lattice-trapped atoms give rise to a frequency shift. Owing to the anti-symmetrization condition of identical fermions, embodied in the Pauli exclusion principle, spin-polarized ^{171}Yb atoms (total atomic angular momentum quantum number $F=1/2$ for the clock states) cannot interact with each other through

Table 1 | Characteristic clock uncertainty budget

Shift	Yb-1 shift	Yb-1 uncertainty	Yb-2 shift	Yb-2 uncertainty	Differential uncertainty
Background gas collisions	−5.5	0.5	−3.6	0.3	0.3
Spin polarization	0	<0.3	0	<0.1	<0.3
Cold collisions ^a	−0.21	0.07	−0.02	0.01	0.07
Doppler	0	<0.02	0	<0.01	0.02
BBR ^a	−2,361.2	0.9	−2,371.7	1.0	0.6
Lattice light (model)	0	0.3	0	0.3	<0.1
Travelling wave contamination	0	<0.1	0	<0.01	<0.1
Lattice light (experimental)	−1.5	0.8	−1.5	0.8	0.2
Second-order Zeeman ^a	−118.1	0.2	−117.9	0.1	0.1
DC Stark	0	<0.07	0	<0.04	<0.08
Probe Stark	0.02	0.01	0.02	0.01	<0.01
Line pulling	0	<0.1	0	<0.1	<0.1
Tunnelling	0	<0.001	0	<0.001	<0.001
Servo error	0.03	0.05	0.03	0.05	<0.01
Optical frequency synthesis	0	<0.1	0	<0.1	<0.1
Total	−2,486.5	1.4	−2,494.7	1.4	
Gravity shift from TT reference surface	180,819	6	180,815	6	0.3
Total shift from TT reference surface	178,333	6	178,320	6	0.8

All values are in units of $10^{-18}\nu_{\text{clock}}$. In the differential measurements, some effects are removed in common-mode between the clocks, while others are uncorrelated, leading to a total uncertainty smaller than either individual clock. Description of each of these effects can be found in the main text and Methods.
^aShifts are calculated and corrected in real time. Listed values represent the average shift for a typical run.

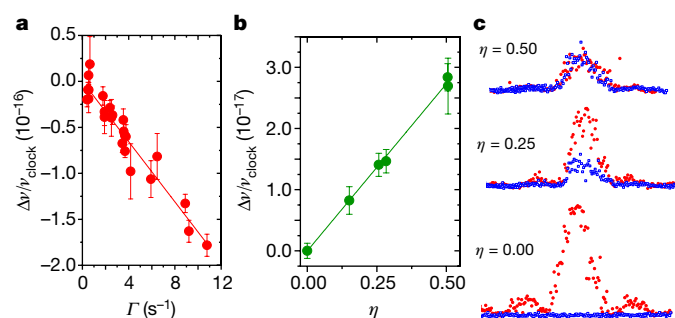


Fig. 2 | Sources of systematic uncertainty. **a**, Measurement of shifts due to collisions with background gas molecules. A fit to the dataset yields a shift of $\Delta\nu/(\nu_{\text{clock}}\Gamma) = -1.64(12) \times 10^{-17}$ s. Error bars are the last point on the total Allan deviation for frequency comparisons (representing half the measurement time of approximately one hour) between Yb-1 and Yb-2. **b**, Cold collisional shift measurements, as a function of spin polarization impurity. A fit to the dataset yields a shift of $\Delta\nu/(\nu_{\text{clock}}\eta) = 5.50(12) \times 10^{-17}$ for the atom number of Yb-1, $N_0 \approx 1,000$ atoms. Error bars are the last point on the total Allan deviation after a measurement time of approximately one hour. **c**, Traces over the Rabi line for variable degrees of optical pumping. Red circles are the π -transition corresponding to the desired spin state, and blue squares are the depleted spin state. From bottom to top, the traces correspond to a detuning of 0 kHz, 680 kHz and 6,800 kHz from the optical pumping transition (natural linewidth is 180 kHz). The value of η (the residual excitation ratio of the depleted spin state) is shown for each trace.

even-partial-wave interactions such as *s*-wave, *d*-wave, *g*-wave, and so on (although we note that indistinguishability may be compromised by inhomogeneous probe excitation; see Methods). We achieve spin polarization by optically pumping on the 556 nm transition in the presence of a 0.45 mT magnetic field²⁹. If spin polarization is incomplete, residual population remains in the depleted spin state, leading to imperfect suppression of *s*-wave collisional shifts. To quantify this effect, we intentionally compromise spin-polarization purity³⁰ by detuning the pumping laser frequency and measure frequency changes between high and low density. As shown in Fig. 2b, we find that the fractional frequency shift for the operational atom number of Yb-1 scales with η , the proportion of the population in the depleted spin state, as $[5.50(12) \times 10^{-17}]\eta$.

The effectiveness of optical pumping into the $m_F = \pm 1/2$ clock state is evaluated by examining the maximal excitation ratio of the π -transition for atoms remaining in the depleted spin state, as displayed in Fig. 2c. For the typical signal-to-noise ratios observed in our experiment, it is straightforward to verify that at least 99.5% of the atoms are prepared in the desired spin state. With this bound we constrain the *s*-wave collisional shift from imperfect spin polarization to be less than 3×10^{-19} for Yb-1. The shift is less than 1×10^{-19} for Yb-2 owing to the larger lattice beam waist (and correspondingly lower atom number density) afforded by the enhancement cavity (see Fig. 1). We note that the spin-1/2 nuclear structure, the simplest of any fermion, provides a consummate advantage to ¹⁷¹Yb in constraining the magnitude of this shift in comparison to other atomic species of interest. Failure to account for imperfect spin polarization can easily lead to errors above the 10^{-18} level.

Benefitting from the careful control of systematic shifts that compromise stability at long timescales, we demonstrate a clock instability that reaches into the 10^{-19} decade, as shown in Fig. 3. This level of performance is demonstrated in two complementary configurations: the blue dataset uses 560-ms Rabi spectroscopy, synchronized between Yb-1 and Yb-2, and the green dataset uses unsynchronized, 510-ms free-evolution-time Ramsey spectroscopy (see Methods). Taken over the course of 72 h, with an uptime of 88%, a single-clock measurement instability of 4.5×10^{-19} is achieved in the former configuration, as determined by the final point of the total Allan deviation at 36 h. An estimated measurement instability of 3.2×10^{-19} is found by fitting

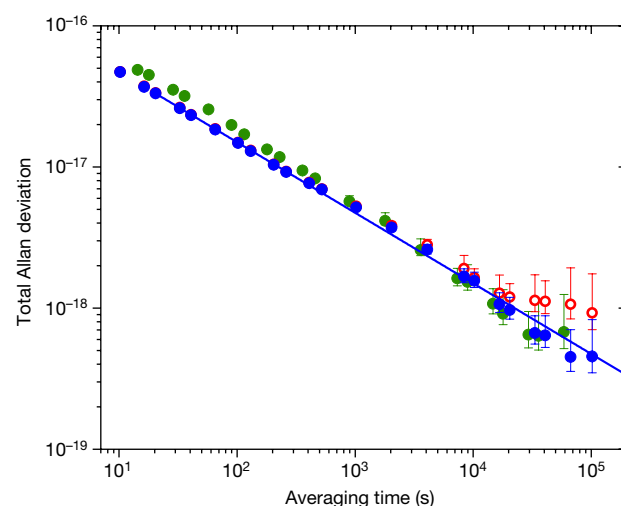


Fig. 3 | Measurement instability. Total Allan deviation (expressed as a fractional frequency) representing an upper bound on the single-clock measurement instability, $(1/\sqrt{2})[\nu_2(t) - \nu_1(t)]/\nu_{\text{clock}}$. The blue and red datasets use synchronized Rabi spectroscopy. The green set uses unsynchronized Ramsey spectroscopy. Blue and green circles are corrected line-by-line for the blackbody shift (see Methods); red circles are uncorrected. The blue line represents the white frequency noise asymptote of $1.5 \times 10^{-16}/\sqrt{\tau}$ for the blue dataset, where τ is the averaging time in seconds. Error bars represent the 1σ uncertainty in the Allan deviation.

a white frequency noise model to the total Allan deviation and extrapolating the fit to the full measurement time. Every 24 h, the clocks were unlocked from the atomic transition so that an evaluation could be performed to ensure full compliance with the uncertainty budget of Table 1. During the clock comparison, the frequency difference is corrected in real time for the blackbody shift (see Methods). The corrected data are consistent with white frequency noise throughout the entire dataset, while a noise floor at 1×10^{-18} is present for the uncorrected dataset. This measurement stability demonstrates the possibility of geopotential determination approaching millimetre-level statistical resolution. Further, this long-time performance does not require synchronized probing schemes, as a comparable instability of 6.4×10^{-19} is obtained with the unsynchronized sequences (Fig. 3).

Making use of the low measurement instability demonstrated, we undertake a campaign of frequency comparisons between the clocks to characterize the reproducibility of the two systems, as shown in Fig. 4. The frequency comparisons are performed with a blinding protocol (see Methods) that prevents the operator from having knowledge of the frequency difference during individual measurements. The results reported here are the culmination of ten blinded measurements, taken over the course of more than a month. A weighted average of these measurements leads to a frequency difference $(\nu_2 - \nu_1)/\nu_{\text{clock}} = [-7 \pm (5)_{\text{stat}} \pm (8)_{\text{sys}}] \times 10^{-19}$ after correcting for all relevant systematic effects, where $\nu_{1(2)}$ is the frequency of Yb-1(2) and $\nu_{\text{clock}} = 518$ THz is the transition frequency. After four early comparisons, not included in the dataset presented here, it was discovered that a faulty wire had removed the grounding connection to the conductive windows of Yb-2, compromising Faraday shielding and leading to a mid- 10^{-18} DC Stark shift between the systems. This experience underscores the indispensability of experimentally investigating reproducibility for substantiating an uncertainty budget.

We note that the gravitational redshift listed in Table 1 is similar for each clock, because the two systems are located in the same laboratory, leading to a small relative uncertainty. However, the redshift transformation to the reference surface of TT has an uncertainty of 6×10^{-18} (limited by the several-centimetre-level state-of-the-art geodetic determination), much larger than the total measurement uncertainty between the clocks³¹. In other words, if these clocks were compared

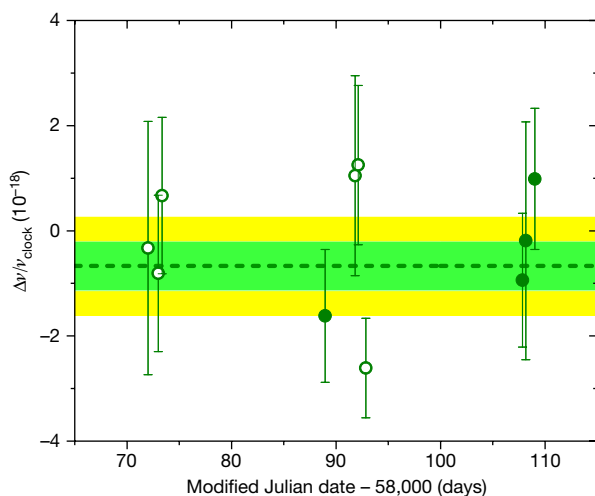


Fig. 4 | Characterization of reproducibility. Green filled (open) circles represent blinded measurements taken in configuration 1 (2), as described in Methods section ‘Experimental setup’. All systematics have been subtracted from the measurements. The dashed line represents the mean, and the green (yellow) shaded region represents the statistical (total) 1σ uncertainty of the measurements: $\Delta\nu/\nu_{\text{clock}} = [-7 \pm (5)_{\text{stat}} \pm (8)_{\text{sys}}] \times 10^{-19}$. The statistical uncertainty is scaled up by the square root of the reduced χ^2 statistic, $\chi^2_{\text{red}} = 1.06$. Error bars represent 1σ uncertainty obtained from the two-clock total Allan deviation, as described in Methods.

across a long baseline or used for remote comparisons with other clocks around the world, the measurement would be limited by gravitational knowledge on Earth’s surface. Refinements of geoid models using satellite-based long-wavelength data³² and terrestrial short-wavelength data³³ may reduce future geoid model uncertainty to some extent, but height uncertainty at or below 1 cm (fractional frequency uncertainties below 1×10^{-18}) remains at best an optimistic goal for conventional geodetic techniques^{21,31}. With performance at the levels demonstrated here, these optical clocks now enable beyond-state-of-the-art geodetic measurements and fundamental physics studies^{7,8,10,13,15}.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0738-2>.

Received: 16 April 2018; Accepted: 20 September 2018;

Published online 28 November 2018.

- Chou, C. W., Hume, D. B., Koelemeij, J. C. J., Wineland, D. J. & Rosenband, T. Frequency comparison of two high-accuracy Al^+ optical clocks. *Phys. Rev. Lett.* **104**, 070802 (2010).
- Nicholson, T. L. et al. Systematic evaluation of an atomic clock at 2×10^{-18} total uncertainty. *Nat. Commun.* **6**, 6896 (2015).
- Huntemann, N., Sanner, C., Lipphardt, B., Tamm, C. & Peik, E. Single-ion atomic clock with 3×10^{-18} systematic uncertainty. *Phys. Rev. Lett.* **116**, 063001 (2016).
- Schioppo, M. et al. Ultra-stable optical clock with two cold-atom ensembles. *Nat. Photon.* **11**, 48–52 (2017).
- Takano, T. et al. Geopotential measurements with synchronously linked optical lattice clocks. *Nat. Photon.* **10**, 662–666 (2016).
- Chou, C. W., Hume, D. B., Rosenband, T. & Wineland, D. J. Optical clocks and relativity. *Science* **329**, 1630–1633 (2010).
- Delva, P. & Lodewyck, J. Atomic clocks: new prospects in metrology and geodesy. *Acta Futura* **7**, 67–78 (2013).
- Lion, G. et al. Determination of a high spatial resolution geopotential model using atomic clock comparisons. *J. Geod.* **91**, 597–611 (2017).
- Grotti, J. et al. Geodesy and metrology with a transportable optical clock. *Nat. Phys.* **14**, 437–441 (2018).
- Bondarescu, R. et al. Ground-based optical atomic clocks as a tool to monitor vertical surface motion. *Geophys. J. Int.* **202**, 1770–1774 (2015).

- Kolkowitz, S. et al. Gravitational wave detection with optical lattice atomic clocks. *Phys. Rev. D* **94**, 124043 (2016).
- Delva, P. et al. Test of special relativity using a fiber network of optical clocks. *Phys. Rev. Lett.* **118**, 221102 (2017).
- Derevianko, A. & Pospelov, M. Hunting for topological dark matter with atomic clocks. *Nat. Phys.* **10**, 933–936 (2014).
- Arvanitaki, A., Huang, J. & Van Tilburg, K. Searching for dilaton dark matter with atomic clocks. *Phys. Rev. D* **91**, 015015 (2015).
- Wcislo, P. et al. Experimental constraint on dark matter detection with optical atomic clocks. *Nat. Astron.* **1**, 0009 (2016).
- Hees, A., Guéna, J., Abgrall, M., Bize, S. & Wolf, P. Searching for an oscillating massive scalar field as a dark matter candidate using atomic hyperfine frequency comparisons. *Phys. Rev. Lett.* **117**, 061301 (2016).
- Roberts, B. M. et al. Search for domain wall dark matter with atomic clocks on board global positioning system satellites. *Nat. Commun.* **8**, 1195 (2017).
- Soffel, M. et al. The IAU 2000 resolutions for astrometry, celestial mechanics, and metrology in the relativistic framework: explanatory supplement. *Astron. J.* **126**, 2687–2706 (2003).
- Vanicek, P., Castle, R. O. & Balazs, E. I. Geodetic leveling and its applications. *Rev. Geophys.* **18**, 505–524 (1980).
- Wang, Y. M., Saleh, J., Li, X. & Roman, D. R. The US Gravimetric Geoid of 2009 (USGG2009): model development and evaluation. *J. Geod.* **86**, 165–180 (2012).
- Denker, H. et al. Geodetic methods to determine the relativistic redshift at the level of 10^{-18} in the context of international timescales: a review and practical results. *J. Geod.* **92**, 487–516 (2018).
- Beloy, K. et al. Atomic clock with 1×10^{-18} room-temperature blackbody Stark uncertainty. *Phys. Rev. Lett.* **113**, 260801 (2014).
- Beloy, K. et al. Faraday-shielded dc Stark-shift-free optical lattice clock. *Phys. Rev. Lett.* **120**, 183201 (2018).
- Brown, R. C. et al. Hyperpolarizability and operational magic wavelength in an optical lattice clock. *Phys. Rev. Lett.* **119**, 253001 (2017).
- Ushijima, I., Takamoto, M., Das, M., Ohkubo, T. & Katori, H. Cryogenic optical lattice clocks. *Nat. Photon.* **9**, 185–189 (2015).
- Akatsuka, T., Takamoto, M. & Katori, K. Optical lattice clocks with non-interacting bosons and fermions. *Nat. Phys.* **4**, 954–959 (2008).
- Campbell, S. L. et al. A Fermi-degenerate three-dimensional optical lattice clock. *Science* **358**, 90–94 (2017).
- Gibble, K. Scattering of cold-atom coherences by hot atoms: frequency shifts from background-gas collisions. *Phys. Rev. Lett.* **110**, 180802 (2013).
- Lemke, N. D. et al. Spin-1/2 optical lattice clock. *Phys. Rev. Lett.* **103**, 063001 (2009).
- Zhang, X. et al. Spectroscopic observation of $\text{SU}(N)$ -symmetric interactions in Sr orbital magnetism. *Science* **345**, 1467–1473 (2014).
- Pavlis, N. K. & Weiss, M. A. A re-evaluation of the relativistic redshift on frequency standards at NIST, Boulder, Colorado, USA. *Metrologia* **54**, 535–548 (2017).
- Bruinsma, S. L. et al. ESA’s satellite-only gravity field model via the direct approach based on all GOCE data. *Geophys. Res. Lett.* **41**, 7508–7514 (2014).
- Smith, D. *The GRAV-D Project: Gravity for the Redefinition of the American Vertical Datum* https://www.ngs.noaa.gov/GRAV-D/pubs/GRAV-D_v2007_12_19.pdf (NOAA, 2007).

Acknowledgements We acknowledge financial support from the National Institute of Standards and Technology, the NASA Fundamental Physics programme, the Defense Advanced Research Projects Agency (DARPA) Quantum Assisted Sensing and Readout (QuASAR) programme and PECASE. R.C.B. acknowledges support from the National Research Council Research Associateship programme. A.D.L. acknowledges support from the International Space Science Institute for contributions to the Spacetime Metrology, Clocks and Relativistic Geodesy Workshop. We also thank T. Fortier and H. Leopardi for femtosecond optical frequency comb measurements, and J. Kitching and D. Hume for careful reading of this manuscript.

Reviewer information Nature thanks K. Bongs, P. Delva and T. Ido for their contribution to the peer review of this work.

Author contributions W.F.M., X.Z., R.J.F., S.A.S., D.N. and A.D.L. carried out the instability and reproducibility measurements. W.F.M., X.Z., S.A.S., K.B., D.N., R.C.B., N.H., G.M., M.S., T.H.Y. and A.D.L. contributed to the evaluation of the uncertainty budget. A.D.L. supervised this work. All authors contributed to the final manuscript. Contributions to this article by workers at NIST, an agency of the US Government, are not subject to US copyright.

Competing interests The authors declare no competing interests.

Additional information

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to A.D.L. **Publisher’s note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Experimental setup. Ytterbium-171 from an effusive cell is slowed and cooled to millikelvin temperature by a three-dimensional magneto-optical trap operating on the allowed $^1S_0 \rightarrow ^1P_1$ transition at 399 nm. The atoms are further cooled to a few microkelvin by a three-stage 556 nm magneto-optical trap on the spin-forbidden $^1S_0 \rightarrow ^3P_1$ transition. The ultracold atomic sample is loaded into an optical lattice close to the magic wavelength, $\lambda_{\text{latt}} = 759$ nm. The lattice of Yb-1, with a $1/e^2$ power radius of 75 μm , is formed by retroreflecting the laser upon itself. A larger radius of 170 μm is achieved in Yb-2 by means of an enhancement cavity²⁴. Both lattices operate at a trap depth of $50 E_r$ (here E_r is the lattice photon recoil energy, given by $E_r = h^2/(2m\lambda_{\text{latt}}^2)$, where h is Planck's constant and m is the mass of ^{171}Yb). The atoms are then cooled to a longitudinal temperature of 500 nK by 20 ms of sideband cooling^{34,35} on the $^1S_0 \rightarrow ^3P_0$ transition at 578 nm, quenched by the $^3P_0 \rightarrow ^3D_1$ transition at 1,388 nm. The atoms are spin-polarized to >99.5% purity on the 556 nm transition. We elect to operate with an atom number, N_0 , of about 1,000 lattice-trapped atoms for both systems.

Rabi spectroscopy with an interrogation time of 560 ms is performed on the ultra-narrow (natural linewidth of 7 mHz), doubly-forbidden 578 nm line, leading to a Fourier-limited full-width at half-maximum of 1.4 Hz, shown in Fig. 1. By applying a bias field of 0.1 mT, the Zeeman spectral lines are split by 400 Hz, and locking is performed by interleaving interrogation between the two lines, as described elsewhere²⁹. The clock laser is stabilized to a 29-cm cavity made of ultra-low expansion glass with an instability assessed as $\leq 1.5 \times 10^{-16}$ by using the atoms as a frequency discriminator⁴. After spectroscopy, the normalized excitation ratio is detected by collecting fluorescence on the $^1S_0 \rightarrow ^1P_1$ transition and repumping on the $^3P_0 \rightarrow ^3D_1$ transition. The total cycle time is 860 ms, leading to a probe time to cycle time ratio of 65%. An atomic shutter blocks the atomic beam during spectroscopy, preventing collisional shifts from the atomic beam as well as excess BBR shifts from the hot oven. The measurements reported here are taken in two complementary configurations: (1) frequency corrections are sent independently to AOM-1 and AOM-2 in Fig. 1, seen by Yb-1 and Yb-2 respectively; and (2) corrections from Yb-1 are sent to AOM-0, seen identically by both systems. In case 1, the frequency difference is simply given by the difference between the two lock integrators. In case 2, the frequency difference is inferred by using Yb-2 as a frequency discriminator. Four measurements, averaging to $-5(6) \times 10^{-19}$, are taken in configuration 1. Six measurements, averaging to $-8(7) \times 10^{-19}$, are taken in configuration 2.

Cold collisional shifts. While s -wave interactions between identical ^{171}Yb atoms are highly suppressed by the Pauli exclusion principle, p -wave collisions are allowed³⁶. However, with $l \geq 1$ angular momentum, a potential barrier forms from the competition of van der Waals attraction and centrifugal repulsion³⁷. As a result, such collisions start to freeze out below the p -wave barrier, calculated³⁸ to be 30 μK . Our trap depth of $50E_r$ corresponds to a maximum atomic temperature of 4.8 μK , and we thus expect collisional shifts to be largely suppressed. Inhomogeneous probe excitation may lead to a residual s -wave collisional shift³⁹, though previous measurements indicate that this effect is usually small compared to p -wave interaction shifts, even at typical ultracold temperatures³⁷. To aid in the measurement of collisional shifts at low trap depths, atom number is enhanced by quenched sideband cooling of atoms trapped in a deeper lattice ($200E_r$), followed by an adiabatic ramp to $50E_r$ depth. Motional sideband spectroscopy is performed, and no statistically significant difference is observed in the motional state distribution between this population and a sideband-cooled population loaded directly into a $50E_r$ lattice. By evaluating the density-dependent collision shift at a range of atom numbers and trap depths, the cold collision shift is measured to be only $-2.1(7) \times 10^{-19}$ for the operational atom number, N_0 , of Yb-1. For Yb-2, the cold collision shift is smaller still, $-0.2(1) \times 10^{-19}$, because of the larger beam waist present in the enhancement cavity. Given the small magnitude of the observed density shift, we note that even operating at an atom number of 40,000, corresponding to a one-second quantum-projection-noise of $< 5 \times 10^{-18}$, leads to a Yb-2 collisional shift uncertainty in the 10^{-19} decade.

Doppler shift. By operating in the Lamb–Dicke regime, the atoms of an optical lattice clock are essentially stationary with respect to the lattice⁴⁰, but motion of the lattice with respect to the laboratory frame results in a first-order Doppler shift. This shift can be explored by introducing a delay before spectroscopy, thus altering the phase and amplitude of the motion experienced during spectroscopy. Uncontrolled, we have observed first-order Doppler shifts in excess of 1×10^{-16} for previous generations of our clock. Lattice motion with respect to the clock laser frame can be reduced by employing clock laser phase-noise cancellation with the lattice retroreflector as a phase reference^{41,42}, though it becomes important to understand how completely this technique can afford suppression. Without cancellation, we evaluate this shift to be $\leq 1 \times 10^{-16}$ for Yb-1 and $\leq 0.5 \times 10^{-16}$ for Yb-2 under the present operating conditions.

To evaluate the fidelity with which the clock laser is coherently transferred to the lattice reference frame, we intentionally induce a large first-order Doppler shift.

This is done by sweeping the voltage of an electro-optic modulator in the clock laser path, leading to an optical path length variation rate ranging from $30 \mu\text{m s}^{-1}$ to $1,500 \mu\text{m s}^{-1}$. After implementing phase-noise cancellation to the lattice retroreflector, we observe that this shift is suppressed by a factor of $\geq 5,000$. Suppression is complete for all optical path length variation rates that were investigated. Applying this suppression factor to the uncanceled Doppler shift, we assess this shift to be no larger than 2×10^{-20} for Yb-1 and 1×10^{-20} for Yb-2.

BBR Stark shift. By far the largest uncanceled shift present in our ^{171}Yb clock is the Stark shift due to BBR from the clock's room-temperature environment. This shift is characterized at better than the part-per-thousand level by means of an in-vacuum thermal shield that provides the atoms with a near-ideal blackbody environment. The temperature of the shield is determined by five platinum resistive thermal detectors: three mounted on the body of the shield, one on the top window and one on a side window. This apparatus has been evaluated comprehensively²², with the following improvements. The resistive thermal detectors are now mounted in a manner that better preserves manufacturer calibration, eliminating post-calibration fidelity as a source of error. The detectors are packed with more aluminium oxide fillings that facilitate heat transfer with the shield in vacuum, reducing self-heating. We have also evaluated inhomogeneous window heating due to lattice laser absorption, finding that for Yb-2 the enhancement cavity increases the effective temperature of the windows on the vertical axis by 440 mK, leading to a shift correction of $6(3) \times 10^{-19}$. For Yb-1, lattice heating is below the mid- 10^{-19} level. With these modifications, the total environmental shift uncertainty is 4×10^{-19} and 5×10^{-19} for Yb-1 and Yb-2, respectively.

During the frequency comparisons, we measure each of the thermal detectors every 100 s and apply a correction to each clock frequency based upon the effective temperature of the system²². We note that the three thermal detectors on the body agree to within the calibration uncertainty of < 5 mK. Even in the tightly controlled thermal environment of the laboratory, real-time corrections are necessary to remove frequency drift between the clocks. In Fig. 3 it is seen that without correcting for the BBR shift, the clocks encounter a noise floor of 1×10^{-18} , consistent with the approximately 100 mK deviations we observe. With BBR corrections properly applied, no noise floor is observed and the clocks average below 5×10^{-19} .

The greater part (8.5×10^{-19}) of the BBR shift uncertainty is due to uncertainty in the coefficient of the so-called dynamic correction to atomic response⁴³. Improved measurement of the 3D_1 lifetime and branching ratio to 3P_0 can reduce this source of uncertainty. Atomic response is common-mode between the two room-temperature clocks, but uncertainty due to blackbody environment is uncorrelated. The differential uncertainty is therefore a quadrature sum of the environmental uncertainty, amounting to 6×10^{-19} .

Lattice light shift. By operating at the magic wavelength, where the electric dipole (E1) polarizabilities of the ground and excited states precisely cancel, clock frequency can be made largely insensitive to lattice laser intensity⁴⁰. This technique allows much better than part-per-million cancellation of the lattice AC Stark shift, but at the current frontier of optical clock performance higher-order polarizabilities due to, for example, M1-, E2- and two-photon E1-transitions are relevant. Combined with motional state quantization in the lattice, these effects introduce a non-polynomial dependence on intensity, as well as dependence on atomic temperature⁴⁴. Recent analysis has found that the scaling of atomic temperature with trap depth can lead to a simplification of the shift, allowing it to be modelled as a polynomial series²⁴. A further implication of these effects is that the E1-polarizability affects higher-order trap-depth-dependent terms, leading to meaningful frequency-dependence of these terms (R.C.B. et al., manuscript in preparation). For the $50E_r$ lattice trap depth we employ here, we find that a cubic fit is sufficient to model relevant light shifts with an error $\leq 3 \times 10^{-19}$. We note that the model uncertainty of the lattice light shift is common-mode between the two clocks and is thus suppressed in the differential uncertainty.

If the forward-going and backward-going lattice beams of Yb-1 do not have the same intensity, a travelling wave is present, leading to an apparent shift of the power series coefficients²⁵. Allowing for an intensity mismatch of up to 15% due to scattering, absorption or imperfect retroreflection and focusing, we conservatively constrain the travelling wave contamination of Yb-1 to be no more than 1×10^{-19} . The enhancement cavity prevents any substantial lattice intensity mismatch from occurring on Yb-2.

To minimize the negative effects of the lattice light shift, we choose to tune our lattice laser to the operational magic wavelength, at which a positive quadratic shift partially cancels a negative linear shift, yielding a shift insensitive to changes in trap depth in the vicinity of our operational trap depth⁴⁴. For our sideband-cooled sample, we find an operational magic wavelength corresponding to a frequency of 394,798,267.7(5) MHz, leading to an experimental lattice light shift uncertainty of 8×10^{-19} . The differential uncertainty is only 2×10^{-19} , as the operational magic wavelength suppresses uncertainty due to small differences in trap depth. The lattice laser is stabilized to a cavity made of ultra-low expansion glass and is measured

on a weekly basis with a resolution better than 10 kHz by an octave-spanning Ti:sapphire optical frequency comb referenced to a calibrated hydrogen maser.

Zeeman shift. As the clock is referenced to a transition between spherically symmetric ($J=0$) electronic states, the atomic wavefunction is insensitive at first order to any non-scalar effect, such as coupling to a magnetic field. Owing to non-zero nuclear spin, a small degree of linear field sensitivity is present. The difference in splitting between π - and σ -transitions can be used to measure the magnetic field, and thus the linear Zeeman sensitivity can be determined⁴⁵. We find that the linear Zeeman effect splits the π -transitions by 199.516(2) Hz per G from centre, in good agreement with a previous measurement²⁹.

By iteratively interrogating the two nuclear spin states and taking the average, the first-order Zeeman shift is cancelled completely. We confirm this experimentally to the 4×10^{-19} level by measuring the clock frequency difference between large magnetic fields with opposite polarity. The measured splitting yields a read-out of the magnetic field, used to precisely determine the second-order Zeeman coefficient. By interleaving between two clocks at different magnetic fields, the coefficient is found to be $-0.06095(7)$ Hz per G², in good agreement with a previous measurement³⁶. Precise spectroscopic determination of the magnetic field is complicated by the existence of a polarization-dependent vector Stark shift from the lattice laser that acts as a pseudo-magnetic field⁴⁵. In practice, it is straightforward to reduce this shift to <100 mHz by using a linearly polarized lattice laser. For Yb-1, the second-order Zeeman shift is $-118.1(2) \times 10^{-18}$, with an uncertainty limited by knowledge of the residual vector Stark shift. For Yb-2, a well-defined linear eigen-polarization is coupled into the build-up cavity, leading to a negligible vector Stark shift²⁴ and a Zeeman uncertainty of only 1×10^{-19} , limited by knowledge of the second-order coefficient. The vector Stark shift is uncorrelated, but the coefficient uncertainty is common-mode, yielding a differential uncertainty of 1×10^{-19} .

DC Stark shift. Stray DC electric fields have been observed to cause Stark shifts of the order of 10^{-13} for some optical lattice clocks⁴⁶. Our thermal shield serves as a Faraday cage to null stray electric fields. During normal clock operation, the body of the shield and the windows, coated with conductive indium-tin oxide, are grounded. By applying voltage to the windows of the thermal shield, the Stark shift due to stray electric fields is found to be consistent with zero at better than the 10^{-19} level on both systems²³.

Probe Stark shift. A measurement of the probe AC Stark shift, arising from the clock light itself, is performed by measuring the clock shift between normal operation and a case where the clock laser is phase-modulated, and the clock transition is excited by a weak resonant sideband 40 dB lower in intensity than the off-resonant carrier. For 560-ms Rabi spectroscopy, the probe shift is found to be 2×10^{-20} , with uncertainty conservatively assessed at 50% of its value. This shift is largely removed in common-mode for the differential measurement.

Line pulling. The presence of other spectroscopic features near the atomic transition can lead to an apparent shift of the line. σ_{\pm} -transitions result from driving the $\Delta m_F = \pm 1$ transition. These transitions, detuned from the clock transition by 1.15 kHz, are typically suppressed to about 1% by using linearly polarized clock light parallel to the quantization axis of the atoms. Imperfect spin polarization might lead to a residual population of clock atoms in the depleted nuclear spin state. This population is suppressed by at least 99.5%, and the π - (σ -) transitions are detuned by ± 400 (∓ 750) Hz. The closest spectroscopic features are the transverse sidebands, detuned by only 60 Hz from the carrier. However, they are highly symmetric due to the large number of transverse motional states and are largely suppressed by probing collinear to the lattice laser. Owing to a relatively narrow Fourier-limited linewidth of 1.4 Hz (Fig. 1), we place a conservative upper bound of 1×10^{-19} on all sources of uncertainty from line pulling.

Tunnelling. In a horizontal lattice, tunnelling of atoms between lattice sites could potentially result in a shift as large as the bandwidth of the Bloch state. By orienting the lattice vertically, adjacent lattice sites are non-degenerate by a frequency Δ_g , given by $h \times \Delta_g = mg(\lambda_{\text{lat}}/2)\cos\theta \approx h \times 1.6$ kHz, where $g = 9.8 \text{ m s}^{-2}$ is the acceleration due to gravity and $\theta \approx 1^\circ$ is the lattice's declination from vertical. This non-degeneracy induces atomic localization through periodic Bloch oscillations, spectroscopically manifested as sidebands corresponding to transitions between Wannier-Stark states. It has been shown theoretically that interference between the carrier and Bloch sidebands can lead to a shift of order of magnitude $\Omega_0\Omega_1/\Delta_g$, where Ω_0 and Ω_1 are respectively the Rabi frequencies of the carrier and the first-order Bloch sideband⁴⁷. By investigating the amplitude of the Bloch sideband for a typical sample at 500 nK (corresponding to a mean longitudinal motional number $\langle n_z \rangle = 0.04$), it is found that $\Omega_1 = (3 \times 10^{-4})\Omega_0$. For 560-ms Rabi spectroscopy, we conservatively assess the tunnelling shift as $<1 \times 10^{-21}$.

Servo error. Local oscillator frequency drift can result in a locked frequency offset from the atomic transition, due in part to the delay between atomic interrogation and frequency feedback correction. We control servo error by probing our transition first in ascending frequency order and then in descending frequency order, cancelling cavity drift at $<500 \mu\text{Hz s}^{-1}$ by digital feed-forward correction

and implementing a second integrator. By analysis of the lock error signal during extended clock operation, we observe servo error consistent with zero at the sub- 10^{-19} level for our longest runs. Owing to the shared local oscillator, servo error is common-mode between the two clocks, and is suppressed to a high degree.

Optical frequency synthesis shifts. Clock operation requires pulsing the 578 nm interrogation laser with an AOM, a process which is known to induce phase chirps. The switching AOM also serves as the phase-noise-cancellation AOM⁴¹ that, when locked, leads to phase transients that are largely random, averaging to <50 mrad, and are suppressed with a time constant of $<20 \mu\text{s}$. Following a previous analysis⁴², we place a conservative upper bound of 1×10^{-20} on this source of uncertainty.

Updating the frequency of a direct digital synthesizer (DDS) can also lead to phase transients that are seen as a frequency shift on the clock transition⁴⁸. By synchronizing a counter to the clock interrogation and directly counting the DDS frequency, we confirm that this technical source of error is consistent with zero at the level of 7×10^{-20} , leading to a differential uncertainty of 1×10^{-19} .

Determination of geopotential. As dictated by relativity, measurements of time in different reference frames should be transformed into a shared reference system. A clock that is elevated above the geoid experiences a gravitational blueshift, as well as a redshift due to increasing centrifugal acceleration⁴⁹. A recent state-of-the-art determination of the geopotential of the Q407 marker on the NIST-Boulder campus found a clock shift of $179,853(6) \times 10^{-18}$ from the reference surface of TT⁵¹. Spirit levelling paired with high-accuracy gravimetry is used to determine a further shift of $810.9(2) \times 10^{-18}$ between the Q407 marker and the laboratory floor. The atomic sample is localized by means of the position of the atomic detection laser, and the heights of the lattice-trapped atoms within Yb-1 and Yb-2 are measured to 2 mm resolution, leading to a further shift of $154.9(2) \times 10^{-18}$ and $151.1(2) \times 10^{-18}$, respectively.

Synchronized operation. The main focus of our study is the characterization of systematic effects and of the level at which they can be controlled. We therefore choose to perform clock comparisons with synchronous interrogation to reduce measurement instability by rejecting the Dick effect in common mode^{50,51}. Despite rejecting the Dick effect, synchronous interrogation does not reject other sources of error, and reaching low instability at long timescales requires control of all systematic shifts. To demonstrate the frequency stability afforded by the clocks, we measure instabilities at the 10^{-19} level for both synchronized and unsynchronized modes of operation, as shown in Fig. 3. For the unsynchronized measurements, we utilize Ramsey spectroscopy with a free-evolution time of 510 ms so that the enhanced quality factor enables a one-second instability similar to the synchronized Rabi case. We note that the one-second instability of 1.5×10^{-16} remains higher than the expected quantum-projection-noise limit ($<5 \times 10^{-17}$, for the conditions discussed above), owing to technical sources of noise.

Blinding protocol and selection criteria. Operator bias can be an important source of error in any physical measurement. For example, for frequency comparisons such as those reported in Fig. 4, it would be possible for the operator to stop the experiment when statistical fluctuations temporarily push the average frequency difference to zero. To eliminate this source of bias, we employ a blinding protocol that adds a large offset to the frequency difference observed by the operator. This offset is pseudorandomly chosen from a uniform distribution spanning ± 1 kHz, many orders of magnitude greater than the sub-millihertz resolution of the experiment. The operator is completely blind to the frequency difference until the termination of the measurement.

Before commencing each blinded measurement, both clocks are comprehensively assessed to verify that they are fully compliant with the uncertainty budget of Table 1. This assessment consists of a checklist of all parameters that can change from day to day. For instance, the operator verifies via motional sideband spectroscopy that the trap depth is within 10% of $50 E_r$ and that the longitudinal temperature is lower than $1 \mu\text{K}$. A post-selection criterion, that data averages as white frequency noise for averaging times greater than 100 s, is also established. An eleventh dataset, with a noise floor of 7×10^{-18} at $\geq 10^4$ s, is excluded because of the latter criterion. We note that, owing to the larger error bars associated with it, inclusion of this set would not meaningfully change the average. With no omissions, each of the ten blinded measurements subject to these criteria is included in the computation of the average frequency difference of $-7(5)_{\text{stat}} \times 10^{-19}$. Statistical error bars are assigned by applying a $1/\sqrt{\tau}$ fit to the total Allan deviation assuming a servo attack time of 20 s and extrapolating to the full measurement time.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

- Curtis, E. A., Oates, C. W. & Hollberg, L. Quenched narrow-line second- and third-stage laser cooling of ^{40}Ca . *J. Opt. Soc. Am. B* **20**, 977–984 (2003).
- Nemitz, N. et al. Frequency ratio of Yb and Sr clocks with 5×10^{-17} uncertainty at 150 seconds averaging time. *Nat. Photon.* **10**, 258–261 (2016).

36. Lemke, N. D. et al. *p*-wave cold collisions in an optical lattice clock. *Phys. Rev. Lett.* **107**, 103902 (2011).
37. Julienne, P. S. & Mies, F. H. Collisions of ultracold trapped atoms. *J. Opt. Soc. Am. B* **6**, 2257–2269 (1989).
38. Dzuba, V. A. & Derevianko, A. Dynamic polarizabilities and related properties of clock states of the ytterbium atom. *J. Phys. B* **43**, 074011 (2010).
39. Swallows, M. D. et al. Suppression of collisional shifts in a strongly interacting lattice clock. *Science* **331**, 1043–1046 (2011).
40. Katori, H., Takamoto, M., Pal'chikov, V. G. & Ovsiannikov, V. D. Ultrastable optical clock with neutral atoms in an engineered light shift trap. *Phys. Rev. Lett.* **91**, 173005 (2003).
41. Ma, L., Jungner, P., Ye, J. & Hall, J. L. Delivering the same optical frequency at two places: accurate cancellation of phase noise introduced by an optical fiber or other time-varying path. *Opt. Lett.* **19**, 1777–1779 (1994).
42. Falke, S., Misera, M., Sterr, U. & Lisdat, C. Delivering pulsed and phase stable light to atoms of an optical clock. *Appl. Phys. B* **107**, 301–311 (2012).
43. Porsev, S. G. & Derevianko, A. Multipolar theory of blackbody radiation shift of atomic energy levels and its implications for optical lattice clocks. *Phys. Rev. A* **74**, 020502 (2006).
44. Katori, H., Ovsiannikov, V. D., Marmo, S. I. & Palchikov, V. G. Strategies for reducing the light shift in atomic clocks. *Phys. Rev. A* **91**, 052503 (2015).
45. Boyd, M. et al. Nuclear spin effects in optical lattice clocks. *Phys. Rev. A* **76**, 022510 (2007).
46. Lodewyck, J., Zawada, M., Lorini, L., Gurov, M. & Lemonde, P. Observation and cancellation of a perturbing dc Stark shift in strontium optical lattice clocks. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **59**, 411–415 (2012).
47. Lemonde, P. & Wolf, P. Optical lattice clock with atoms confined in a shallow trap. *Phys. Rev. A* **72**, 033409 (2005).
48. Lee, W. D., Shirley, J. H., Walls, F. L. & Drullinger, R. E. Systematic errors in cesium beam frequency standards introduced by digital control of the microwave excitation. *Proc. IEEE Int. Freq. Control Symp. Expo.* 113–117 (1995).
49. Hofmann-Wellenhof, B. & Moritz, H. *Physical Geodesy* (Springer, Vienna, 2005).
50. Bize, S. et al. Interrogation oscillator noise rejection in the comparison of atomic fountains. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **47**, 1253–1255 (2000).
51. Takamoto, M., Takano, T. & Katori, H. Frequency comparison of optical lattice clocks beyond the Dick limit. *Nat. Photon.* **5**, 288–292 (2011).

Extreme-ultraviolet refractive optics

L. Drescher¹, O. Kornilov^{1*}, T. Witting¹, G. Reitsma¹, N. Monserud¹, A. Rouzée¹, J. Mikosch¹, M. J. J. Vrakking¹ & B. Schütte^{1*}

Refraction is a well-known optical phenomenon that alters the direction of light waves propagating through matter. Microscopes, lenses and prisms based on refraction are indispensable tools for controlling light beams at visible, infrared, ultraviolet and X-ray wavelengths¹. In the past few decades, a range of extreme-ultraviolet and soft-X-ray sources has been developed in laboratory environments^{2–4} and at large-scale facilities^{5,6}. But the strong absorption of extreme-ultraviolet radiation in matter hinders the development of refractive lenses and prisms in this spectral region, for which reflective mirrors and diffractive Fresnel zone plates⁷ are instead used for focusing. Here we demonstrate control over the refraction of extreme-ultraviolet radiation by using a gas jet with a density gradient across the profile of the extreme-ultraviolet beam. We produce a gas-phase prism that leads to a frequency-dependent deflection of the beam. The strong deflection near to atomic resonances is further used to develop a deformable refractive lens for extreme-ultraviolet radiation, with low absorption and a focal length that can be tuned by varying the gas pressure. Our results open up a route towards the transfer of refraction-based techniques, which are well established in other spectral regions, to the extreme-ultraviolet domain.

Refraction of light is omnipresent in nature, where it forms the basis for the functionality of the human eye and the observation of a rainbow. It is exploited in many applications in the visible, infrared and ultraviolet spectral regions. For instance, refractive errors of the eye are corrected by glasses or contact lenses, and optical microscopes enable the magnification of small objects or structures. In the context of laser physics, refractive lenses are extensively used to focus or (de-)magnify laser beams. Dispersion and deflection of light by optical prisms is used to compress or stretch ultrashort laser pulses.

When Röntgen discovered X-rays in 1895, he attempted refraction experiments using prisms and lenses⁸. Because he observed no significant deflection of the X-rays, he concluded that refractive lenses were not suitable for focusing X-ray radiation. A century later, a compound refractive lens consisting of a lens array was nevertheless developed for the hard X-ray regime, assisted by the comparably low absorption in this spectral region. Compound refractive lenses are used to focus X-rays emitted from modern synchrotron⁹ and free-electron laser facilities^{10,11}. They have been applied for hard X-ray microscopy¹², for X-ray nanofocusing¹³ and for the investigation of crystal scattering¹⁴, as well as for coherent diffractive imaging of nanoscale samples¹⁵.

Refractive elements have so far been missing in the extreme-ultraviolet (XUV) range but are highly desirable. For instance, refractive lenses could be used to focus XUV pulses without changing the propagation direction, thereby providing considerable flexibility. The use of specially designed microscopic refractive lenses has been proposed^{16,17}. However, the need to use very thin lenses with a sophisticated design, owing to the strong absorption of XUV radiation, makes practical implementation challenging.

Here, we demonstrate that control over the refraction of XUV pulses can be achieved by using gases instead of solids. We exploit the fact that close to atomic resonances, the refractive index n has a dispersive lineshape, as depicted in the top part of Fig. 1a. As the photon energy approaches the resonant energy, n first increases and then steeply decreases across the resonance to values below unity, before increasing again.

Our scheme for control over the refraction of XUV pulses with an inhomogeneous gas target is presented in the middle and bottom panels of Fig. 1a. The XUV pulses pass through a gas jet, which propagates in a direction perpendicular to the XUV beam and has a density gradient in the vertical direction (middle panel in Fig. 1a). When the XUV pulse crosses the gas jet off-centre, the jet acts as a prism and induces angular dispersion and deflection of the XUV radiation. For an XUV beam that is incident below the centre of the gas jet, spectral components of the beam for which $n > 1$ are deflected upwards (red colour in the bottom panel of Fig. 1a), whereas spectral components of the beam for which $n < 1$ are deflected downwards (blue colour in the bottom panel of Fig. 1a).

An experimental demonstration of this concept is presented in Fig. 1b, c. Figure 1b shows an XUV spectrum produced by high-harmonic generation (HHG) using near-infrared (NIR) pulses with a duration of 4.5 fs. The spectrum was measured on a 2D detector, in which the horizontal axis represents the axis along which the XUV spectrum is dispersed using a flat-field grating (see Methods). When the broadband HHG pulses propagate 0.3 mm below the centre of a dense He gas jet, the XUV spectrum is strongly modified (see Fig. 1c). Spectral components below the $1s\ np$ ($n = 2, 3, \dots$) resonances of He are deflected upwards, whereas spectral components above these resonances are deflected downwards.

Microscopically, refraction is explained in terms of oscillating electric dipoles induced by the XUV pulse. The incoming XUV pulse excites atoms that re-emit radiation at the same photon energy (free induction decay)^{18–20}. Our prism uses the fact that this re-emitted radiation is phase-shifted with respect to the exciting XUV pulse. Because of the induced gas density gradient, the upper part of the XUV pulse acquires a different phase shift from the lower part, and this leads to a tilt of the XUV wavefront. The wavefront tilt depends on the refractive index and therefore increases close to a resonance²¹. For example, at a photon energy of 21 eV, that is, 0.22 eV below the $1s\ 2p$ resonance, the wavefront tilt is 0.07° at the gas density used in the present experiment.

We have simulated refraction in a gas jet using the Lorentz–Lorentz formula, assuming an XUV beam with a Gaussian spatial profile and using the properties of the $1s\ np$ absorption series of He (see Methods). The phase accumulated by the XUV pulse during propagation through the gas medium is calculated using the eikonal approximation. Propagation of the XUV beam in free space is calculated using the small-angle approximation to the Kirchhoff’s diffraction formula solved by the Fourier transform method (see Methods for details). As shown in Fig. 1d, the simulation reproduces the experimental result well.

The deflection of the XUV beam can be controlled by varying the gas pressure. Angle-resolved spectra for He backing pressures of 1 bar, 3 bar and 9 bar are presented in Fig. 2a–c and show increasing deflection for increasing backing pressure. The average deflection angle as a function of the photon energy (determined by comparing the centre-of-mass of the distribution along the vertical axis with and without the He gas jet) is plotted in Fig. 2d for backing pressures of 3 bar (cyan solid curve) and 9 bar (orange solid curve). For small angles, the deflection is proportional to the refractivity (that is, $n - 1$), which was calculated using the Lorentz–Lorentz formula (see Methods). The shapes of the measured deflection angles (solid curves) and the calculated refractivities (dotted curves) agree well, apart from the region near resonance,

¹Max-Born-Institut, Berlin, Germany. *e-mail: kornilov@mbi-berlin.de; schuette@mbi-berlin.de

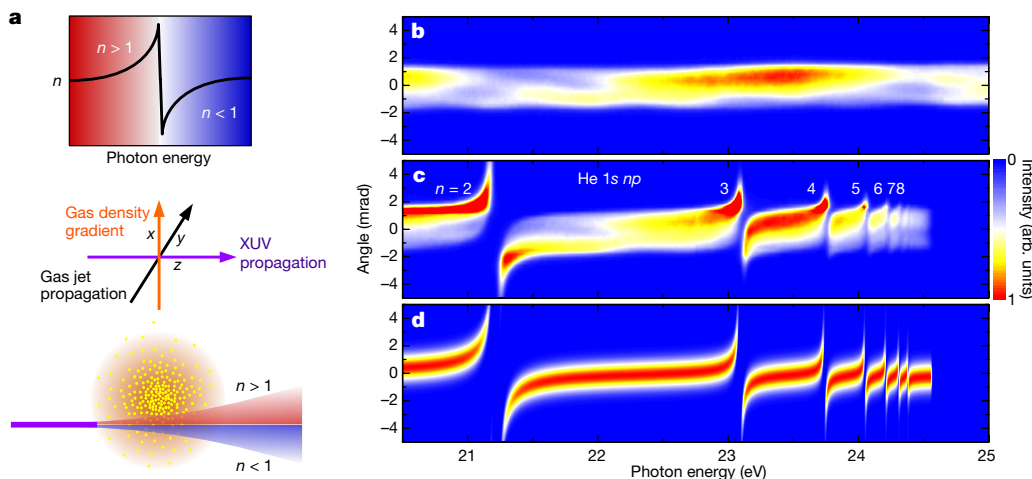


Fig. 1 | XUV refractive prism. **a**, Top, dispersive lineshape of the refractive index in the vicinity of an atomic resonance. Spectral components at photon energies below the resonance ($n > 1$) are indicated in red, components at energies above the resonance ($n < 1$) in blue. Middle, experimental configuration, showing an XUV pulse (violet arrow) that crosses a gas jet (black arrow), which has a density gradient in the vertical direction (orange arrow), at right angles. Bottom, deflection of an XUV pulse propagating below the centre of the gas jet. **b**, Angle-resolved spectrum of a broadband HHG pulse measured in the absence of the gas jet. The angular divergence of the XUV beam in the vertical direction is reflected in the spatial distribution along the vertical axis. arb. units, arbitrary units. **c**, The same spectrum after propagation at a distance of 0.3 mm below the centre of a dense He gas jet (generated using a backing pressure of 10 bar) shows clear signatures of refraction.

where the angular acceptance and the resolution of the XUV spectrometer are no longer sufficient.

The observed deflection of XUV radiation in the vicinity of atomic resonances can be exploited for the design of an XUV refractive lens. In a second set of experiments, high harmonics with a narrow bandwidth of 100–200 meV and a low beam divergence were generated using a 12-m-long beamline²² (see Methods). Figure 3a depicts the spatially resolved spectrum around 20.2 eV (corresponding to the 13th harmonic) as recorded at a distance of 6 m behind the HHG cell. The photon energy of 20.2 eV is about 1 eV below the $1s\ 2p$ resonance of He. The spatial extension of the harmonic along the vertical axis (2.7 mm; see Fig. 3d) corresponds to a full-width at half-maximum (FWHM) divergence of 0.45 mrad. When a He gas jet with a parabolic profile²³ and a spatial extent of about 2.5 mm (that is, similar to the XUV beam diameter, which is 2.3 mm at this point) is placed 0.9 m in front of the detector, the former acts as a lens, as sketched in the inset of Fig. 3b. Figure 3b,c demonstrates focusing of radiation at 20.2 eV for two different backing pressures. Figure 3d shows that the FWHM in the vertical direction is reduced from 2.7 mm to 410 μm by operating the gas jet at a backing pressure of 12 bar (the highest pressure used in the experiment, leading to a peak density in the experiment of about 1×10^{20} atoms cm^{-3}). We found that absorption of the XUV beam by the He lens is small: that is, below the estimated detection threshold of 5%. The geometry of the current experiment leads to focusing in one dimension, analogous to focusing by a cylindrical lens. A sequence of two perpendicularly placed gas jets, each with a cylindrically shaped density gradient, could be used to focus XUV pulses both horizontally and vertically.

As the deflection of XUV radiation increases for photon energies approaching an atomic resonance, we have studied another example using radiation at 14.0 eV (corresponding to the ninth harmonic; see Fig. 3e), which is close to the $3p^5\ 5s$ (at 14.09 eV) and $3p^5\ 3d$ (at 14.15 eV) resonances of Ar. In this case, an Ar gas jet with a moderate backing pressure of 2.5 bar (corresponding to a peak density in the interaction region of about 2×10^{19} atoms cm^{-3}) was used to focus the XUV radiation, as shown in Fig. 3f. On further increasing the gas

Spectral components with photon energies below the $1s\ np$ resonances of He are deflected upwards, whereas spectral components above these resonances are deflected downwards. The deflection angles are largest close to the $1s\ 2p$ resonance and decrease for higher resonances, owing to the decreasing oscillator strengths. Above the ionization potential of He (at 24.58 eV), the XUV radiation is strongly absorbed. Owing to ageing effects, the sensitivity of the detector was reduced in regions where the undisturbed HHG spectrum is recorded (as in **b**) compared with regions where the deflected XUV radiation is observed. This makes the deflected XUV radiation appear more intense. **d**, Simulation of the XUV refraction in an inhomogeneous He gas jet, taking into account $1s\ np$ resonances with $n = 2, 3, \dots, 8$. The simulation indicates that for a backing pressure of 10 bar, a gas jet with a peak density of 9×10^{19} atoms cm^{-3} (corresponding to a pressure of 3.7 bar at 300 K) was achieved in the interaction zone.

backing pressure to 4 bar, the beam size at the detector increased again (Fig. 3g). In this case, the focal plane shifts closer to the jet, and a divergent beam is detected.

A minimum beam size of 270 μm was observed in the experiments with the Ar lens (Fig. 3h), which is small enough for many applications including photoion and photoelectron spectroscopy. Some applications, however, such as the investigation of XUV-induced nonlinear processes^{22,24–27} and single-shot HHG-driven coherent diffractive imaging using photon energies around 20 eV (ref. ²⁸) require substantially smaller XUV spot sizes. The achievable focal spot size is limited by geometric and chromatic aberrations. For ideal focusing conditions, the profile of the gas density integrated along the XUV beam propagation axis needs to be parabolic. Although the gas density profile generated by a cylindrical nozzle is parabolic to a good approximation²³, deviations from the parabolic shape lead to geometric aberrations, affecting the focal spot size that can be achieved. Furthermore, a density gradient is present along the propagation axis of the gas beam that also leads to geometric aberration. In the future, the gas density profile may be optimized by tailoring the gas nozzle designs²³.

Assuming a parabolic gas density profile, we have simulated the spot sizes achieved by an Ar lens for a collimated XUV beam at 14.0 eV with a FWHM diameter of 1.9 mm. The XUV spot size at a distance of 90 cm behind the gas lens depends on the photon energy, as shown in Fig. 4a. This chromatic aberration, which is a direct consequence of the variation of the refractive index within the XUV bandwidth, results in a spot size that is larger than that of a monochromatic XUV beam. Note that this effect is not visible in the experimental data owing to the spectral resolution of about 100 meV and the spatial resolution of about 100 μm . The gas-density-dependent spot size at a distance of 90 cm from an Ar lens is plotted in Fig. 4b, showing a minimum spot size of 74 μm for an XUV pulse with a bandwidth of 160 meV, which is similar to the bandwidth of the ninth harmonic observed in the experiment (black curve in Fig. 4b). The measured spot sizes show little variation over a broad range of gas densities because of chromatic aberration, which explains the behaviour shown in the inset of Fig. 3h.

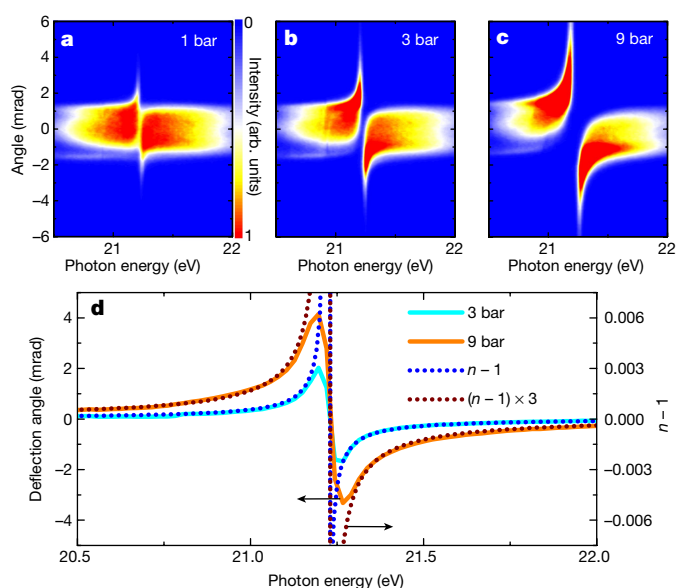


Fig. 2 | Control over XUV deflection by gas pressure. **a–c**, Angle-resolved XUV spectra after propagation at a distance of 0.3 mm below the centre of a He gas jet, for backing pressures of 1 bar (**a**), 3 bar (**b**) and 9 bar (**c**). **d**, The average deflection angle as a function of the photon energy for backing pressures of 3 bar (corresponding to a peak pressure in the interaction zone of about 1 bar; cyan solid curve) and 9 bar (orange solid curve). Here the vertical scale on the left axis applies, as indicated by the upper arrow. For comparison, the calculated refractivity (that is, $n - 1$) at standard temperature (273.15 K) and standard pressure (1 bar) is plotted on top of the deflection results (blue dotted curve). The vertical scale on the right axis applies, as indicated by the lower arrow. Note that the calculated refractivity is proportional to the pressure. The brown dotted curve shows the calculated refractivity multiplied by a factor of 3.

When reducing the XUV bandwidth in the calculation to 2 meV, a minimum spot size of 40 μm is obtained (red curve in Fig. 4b). The chromatic aberration is reduced when exploiting refraction due to a

resonance that is further away from the XUV photon energy, as shown for a He lens with a FWHM diameter of 2.4 mm in Fig. 4c. A minimum spot size of 28 μm is obtained in this case using a pulse with a bandwidth of 240 meV (similar to the bandwidth of the 13th harmonic used in the experiment shown in Fig. 3; black curve in Fig. 4d), and it is reduced to 20 μm for a pulse with a bandwidth of 2 meV (red curve in Fig. 4d). Our simulations show that by further increasing the gas density to $3.1 \times 10^{20} \text{ atoms cm}^{-3}$, corresponding to a shorter focal length of 30 cm, a focal spot size below 10 μm could be achieved (resulting in an XUV intensity of up to $10^{13} \text{ W cm}^{-2}$). This is in the range of recent experiments studying XUV-induced Rabi cycling²⁹, XUV double ionization of atoms²⁶ and single-shot coherent diffractive imaging²⁸, where spot sizes between 3 μm and 16 μm were used, thus putting us in the realm where XUV nonlinear optics experiments become possible.

When using a refractive lens to focus ultrashort XUV pulses, another important aspect is the XUV pulse duration at the focus. Because HHG and free-electron laser pulses have an intrinsic negative chirp^{30,31}, a refractive lens, which induces a positive chirp, can lead to compression of the XUV pulses. Assuming a pulse with a duration of 24 fs and a chirp of -8 meV fs^{-1} , which is in the range of previous measurements for the 13th harmonic³², our simulations show compression to 16 fs by a He lens with a peak gas density of $4.9 \times 10^{19} \text{ atoms cm}^{-3}$ (corresponding to a focal length of 1.9 m). Note that this value is larger than the Fourier-limited pulse duration of 8 fs owing to the nonlinear chirp that is introduced by the lens. When increasing the peak gas density to $1.1 \times 10^{20} \text{ atoms cm}^{-3}$ (corresponding to a focal length of 90 cm), we predict a moderate stretching from 24 fs to 29 fs. Focusing of shorter XUV pulses may be achieved by combining a refractive lens with another focusing element. For example, the development of a multi-component lens consisting of an XUV refractive lens and a Fresnel zone plate was suggested^{16,17}. It was theoretically shown that these multi-component lenses can be used to focus broadband attosecond pulses to nanometre spot sizes¹⁷, which may enable the investigation of electronic processes with attosecond temporal and nanometre spatial resolution.

In conclusion, we have presented a method to deflect and focus XUV pulses by using the inhomogeneity of a gas jet placed in the way of

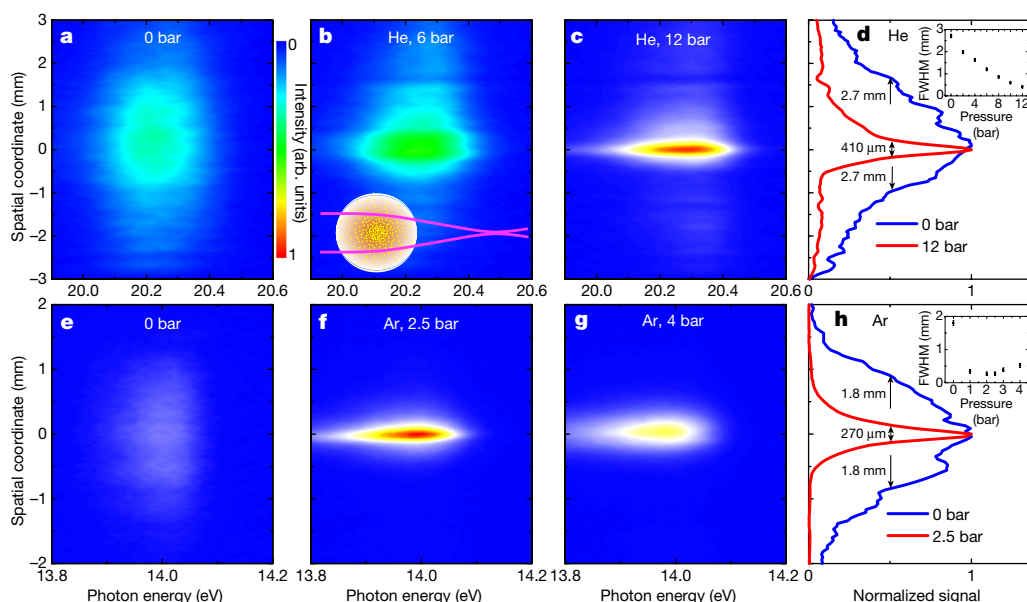


Fig. 3 | XUV refractive lens. **a**, Spatially resolved spectrum of unfocused XUV radiation at 20.2 eV (corresponding to the 13th harmonic). **b, c**, The divergence of this harmonic is altered after propagation through a He gas jet (see inset of **d**), as shown for backing pressures of 6 bar (**b**) and 12 bar (**c**). **d**, Comparison of the vertical beam profiles using backing pressures of 0 bar (blue curve) and 12 bar (red curve). The inset shows the pressure-dependent spot size, where the error bars reflect the uncertainties in determining the spot sizes. **e**, Spatially resolved spectrum of radiation

at 14.0 eV (corresponding to the ninth harmonic), which is close to the 3d and 5s resonances of Ar. **f**, Focusing of this harmonic is achieved by an Ar gas jet at a backing pressure of 2.5 bar. **g**, When further increasing the backing pressure to 4 bar, an increasing beam size is observed, because the Ar lens focuses the XUV beam between the gas jet and the detector. **h**, The vertical beam profiles for Ar backing pressures of 0 bar (blue curve) and 2.5 bar (red curve). The inset shows the pressure-dependent spot size, where the error bars reflect the uncertainties in determining the spot sizes.

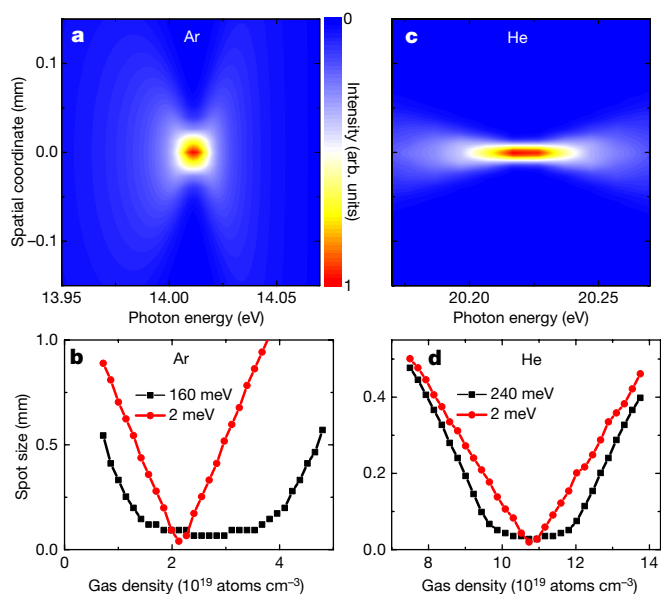


Fig. 4 | Simulation of the XUV focus. **a**, Simulated focus in the vertical direction as a function of the photon energy following propagation of an XUV pulse at 14.0 eV (1.9 mm FWHM diameter) through an Ar gas jet with a peak density of 2.2×10^{19} atoms cm^{-3} (corresponding to a pressure of 0.9 bar at 300 K). Because of chromatic aberration, the XUV spot size depends on the photon energy. **b**, Spot size as a function of the Ar gas density for XUV pulses with a bandwidth of 160 meV (black curve) and 2 meV (red curve), showing minimal spot sizes of 74 μm and 40 μm , respectively. **c**, The chromatic aberration is reduced for photon energies that are further away from the resonance. This is shown for the example of an XUV pulse at 20.2 eV (2.4 mm FWHM diameter) that propagates through a He gas jet with a peak density of 1.1×10^{20} atoms cm^{-3} (corresponding to a pressure of 4.3 bar at 300 K). **d**, Spot size as a function of He gas density for XUV pulses with a bandwidth of 240 meV (black curve) and 2 meV (red curve), which exhibit minimal spot sizes of 28 μm and 20 μm .

an XUV beam. Our results enable the transfer of concepts based on refractive optics that are widely used in other spectral regions to the XUV regime, including microscopy, nanofocusing and the compression of ultrashort pulses. XUV gas-based lenses have several advantages, including their high transmission, deformability and tunability (by varying the gas composition, the gas pressure and the gas jet geometry). Compared with reflective mirrors that are often used to focus XUV pulses, these XUV lenses are immune to damage (because the gas sample is constantly replenished) and preserve the propagation direction of the incoming XUV light, thereby aiding their use in experimental setups. Refractive XUV lenses may also be used in combination with other optical elements and techniques, such as the recently demonstrated spectral selection of harmonics using spatial filtering³³.

Refractive XUV gas-phase lenses can be designed for photon energies between 10 eV and 24 eV by carefully selecting appropriate atoms or molecules for different photon energies. In the future, this range might be extended to higher photon energies by developing lenses that exploit refraction in an inhomogeneous plasma consisting of highly charged ions and electrons.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0737-3>.

Received: 27 April 2018; Accepted: 10 September 2018;
Published online 28 November 2018.

- Snigirev, A., Kohn, V., Snigireva, I. & Lengeler, B. A compound refractive lens for focusing high-energy X-rays. *Nature* **384**, 49–51 (1996).
- Ferray, M. et al. Multiple-harmonic conversion of 1064 nm radiation in rare gases. *J. Phys. B* **21**, L31 (1988).

- Rocca, J. J. Table-top soft X-ray lasers. *Rev. Sci. Instrum.* **70**, 3799–3827 (1999).
- Giulietti, D. & Gizzi, L. A. X-ray emission from laser-produced plasmas. *Riv. Nuovo Cim.* **21**, 1–93 (1998).
- Marr, G. V. *Handbook on Synchrotron Radiation: Vacuum Ultraviolet and Soft X-ray Processes* Vol. 2 (Elsevier, Amsterdam, 2013).
- Allaria, E. et al. Highly coherent and stable pulses from the Fermi seeded free-electron laser in the extreme ultraviolet. *Nat. Photon.* **6**, 699–704 (2012).
- Baez, A. V. A self-supporting metal Fresnel zone-plate to focus extreme ultra-violet and soft X-rays. *Nature* **186**, 958 (1960).
- Röntgen, W. C. Über eine neue Art von Strahlen: Vorläufige Mittheilung. *Sitzungsber. Phys. Med. Gesell. Würzburg* (1895).
- Santoro, G. et al. Use of intermediate focus for grazing incidence small and wide angle X-ray scattering experiments at the beamline P03 of PETRA III, DESY. *Rev. Sci. Instrum.* **85**, 043901 (2014).
- Chollet, M. et al. The X-ray pump-probe instrument at the Linac Coherent Light Source. *J. Synchrotron Radiat.* **22**, 503–507 (2015).
- Heimann, P. et al. Compound refractive lenses as pre-focusing optics for X-ray FEL radiation. *J. Synchrotron Radiat.* **23**, 425–429 (2016).
- Lengeler, B. et al. A microscope for hard X-rays based on parabolic compound refractive lenses. *Appl. Phys. Lett.* **74**, 3924–3926 (1999).
- Schroer, C. G. et al. Hard X-ray nanoprobe based on refractive X-ray lenses. *Appl. Phys. Lett.* **87**, 124103 (2005).
- Meijer, J.-M. et al. Observation of solid–solid transitions in 3D crystals of colloidal superballs. *Nat. Commun.* **8**, 14352 (2017).
- Schroer, C. G. et al. Coherent X-ray diffraction imaging with nanofocused illumination. *Phys. Rev. Lett.* **101**, 090801 (2008).
- Wang, Y., Yun, W. & Jacobsen, C. Achromatic Fresnel optics for wideband extreme-ultraviolet and X-ray imaging. *Nature* **424**, 50 (2003).
- Pan, H. et al. Low chromatic Fresnel lens for broadband attosecond XUV pulse applications. *Opt. Express* **24**, 16788–16798 (2016).
- Hahn, E. L. Nuclear induction due to free Larmor precession. *Phys. Rev.* **77**, 297–298 (1950).
- Wu, M., Chen, S., Camp, S., Schafer, K. J. & Gaarde, M. B. Theory of strong-field attosecond transient absorption. *J. Phys. B* **49**, 062003 (2016).
- Bengtsson, S. et al. Space-time control of free induction decay in the extreme ultraviolet. *Nat. Photon.* **11**, 252–258 (2017).
- Liao, C.-T., Sandhu, A., Camp, S., Schafer, K. J. & Gaarde, M. B. Beyond the single-atom response in absorption line shapes: probing a dense, laser-dressed helium gas with attosecond pulse trains. *Phys. Rev. Lett.* **114**, 143002 (2015).
- Schütte, B., Arbeiter, M., Fennel, T., Vrakking, M. J. J. & Rouzée, A. Rare-gas clusters in intense extreme-ultraviolet pulses from a high-order harmonic source. *Phys. Rev. Lett.* **112**, 073003 (2014).
- Semushin, S. & Malka, V. High density gas jet nozzle design for laser target production. *Rev. Sci. Instrum.* **72**, 2961–2965 (2001).
- Tzallas, P., Charalambidis, D., Papadogiannis, N. A., Witte, K. & Tsakiris, G. D. Direct observation of attosecond light bunching. *Nature* **426**, 267 (2003).
- Takahashi, E. J., Lan, P., Mücke, O. D., Nabekawa, Y. & Midorikawa, K. Attosecond nonlinear optics using gigawatt-scale isolated attosecond pulses. *Nat. Commun.* **4**, 2691 (2013).
- Manschuetus, B. et al. Two-photon double ionization of neon using an intense attosecond pulse train. *Phys. Rev. A* **93**, 061402 (2016).
- Barillot, T. R. et al. Towards XUV pump-probe experiments in the femtosecond to sub-femtosecond regime: new measurement of the helium two-photon ionization cross-section. *Chem. Phys. Lett.* **683**, 38–42 (2017).
- Rupp, D. et al. Coherent diffractive imaging of single helium nanodroplets with a high harmonic generation source. *Nat. Commun.* **8**, 493 (2017).
- Flögel, M. et al. Rabi oscillations in extreme ultraviolet ionization of atomic argon. *Phys. Rev. A* **95**, 021401 (2017).
- Schafer, K. J. & Kulander, K. C. High harmonic generation from ultrafast pump lasers. *Phys. Rev. Lett.* **78**, 638–641 (1997).
- Frühling, U. et al. Single-shot terahertz-field-driven X-ray streak camera. *Nat. Photon.* **3**, 523 (2009).
- Mauritsson, J. et al. Measurement and control of the frequency chirp rate of high-order harmonic pulses. *Phys. Rev. A* **70**, 021801 (2004).
- Valentin, C. et al. Spectral selection of high harmonics via spatial filtering. In *High-Brightness Sources and Light-driven Interactions* HW3A.3 (Optical Society of America, 2018).

Acknowledgements We thank A. A. Ünal and R. Schumann for their support with the laser systems. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie-Sklodowska-Curie grant agreement no. 641789 MEDEA.

Reviewer information Nature thanks J. Cryan, M. Gaarde and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions L.D. and B.S. performed the prism experiments. B.S. performed the lens experiments. O.K. carried out the simulations. All authors discussed the results and contributed to writing the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to O.K. or B.S. **Publisher's note**: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

XUV prism experiments. The XUV prism experiments were performed at an HHG beamline that was previously described in detail^{34–36}. NIR pulses with a duration of 4.5 fs were obtained by spectrally broadening the output of a commercial Ti:sapphire amplifier in a differentially pumped hollow-core fibre that was filled with Ne. Temporal compression of the broadened spectrum was achieved using chirped mirrors. High harmonics were generated by focusing the compressed NIR pulses into a gas cell that was filled with Xe. A 100-nm-thick Al filter was used to block the NIR beam, and the XUV pulses were refocused by a toroidal mirror. Angle-resolved XUV spectra were recorded by an XUV spectrometer that consisted of a flat-field grating, a multichannel plate/phosphor screen assembly and a digital camera. To generate an XUV spectrum that covers the entire range between 20.5 eV and 25 eV, we used a relatively high NIR intensity for the HHG (approximately 3×10^{14} W cm⁻²). This was previously shown to lead to phase-matching of the contributions from both short and long trajectories, resulting in a complex spectrum and an XUV photon energy-dependent divergence³⁷, as also observed in the spectrum shown in Fig. 1b.

Control over the XUV refraction was achieved by a pulsed gas jet that was positioned near the XUV focus. The gas jet was generated by a piezoelectric valve with a nozzle diameter of 0.5 mm. A three-dimensional manipulator was used to position the gas jet with respect to the XUV focal spot. The XUV beam crossed the gas beam at a distance of about 100 μ m from the exit of the nozzle, where gas densities up to 10^{20} cm⁻³ were achieved. We estimate that the peak pressure in the interaction region is smaller than the applied backing pressure by a factor of about 3 (when assuming a temperature of 300 K). The XUV focal spot size (about 100 μ m) was small compared to the extension of the gas jet at the laser position (about 1 mm). The gas density gradient along the propagation direction of the gas beam is estimated to be much smaller than the density gradient perpendicular to this propagation direction.

XUV lens experiments. The XUV lens experiments were performed at a second HHG beamline, where harmonics with a narrow bandwidth are available²². NIR pulses with a pulse energy of 30 mJ and a duration of 35 fs were generated using a home-built Ti:sapphire amplifier³⁸. High harmonics were generated by focusing the NIR pulses using a spherical mirror with a focal length of 5 m into a 10-cm-long gas cell that was filled with Xe.

To focus the XUV beam, a piezoelectric valve with a nozzle diameter of 1 mm was positioned at a distance of 5 m behind the generation cell. Before propagating through the gas jet, the HHG beam was truncated by a slit to a horizontal width of 200 μ m to increase the spectral resolution in the spectrometer located downstream. A 100-nm-thick Al filter was used in the He experiment, which was removed for the Ar experiment, because it absorbs radiation at 14.0 eV. An XUV spectrometer consisting of a plane grating, a multichannel plate/phosphor screen assembly and a digital camera was used to spectrally and spatially characterize the XUV pulses.

Simulations. To simulate the refraction of XUV pulses following propagation through an inhomogeneous gas jet, we calculated the complex refractive index \tilde{n} using the Lorentz–Lorenz formula³⁹

$$\frac{\tilde{n}^2 - 1}{\tilde{n}^2 + 2} = N(x) \frac{e^2}{3m_e \epsilon_0} \sum_j \frac{f_j}{\omega_{0j}^2 - \omega^2 - i\Gamma_j \omega}$$

Here $N(x)$ is the (spatially dependent) atomic density, e is the electron charge, m_e is the electron mass, ϵ_0 is the vacuum permittivity, f_j is the oscillator strength of the transition j , ω is the angular frequency of the XUV light, and ω_{0j} and Γ_j are the resonant frequencies and widths, respectively. Values for the oscillator strengths and linewidths were taken from refs ^{40,41}. The real part of the refractive index, n , and the absorption index, β , were extracted according to the following equations:

$$n = \sqrt{0.5[|\tilde{n}|^2 + \text{Re}(\tilde{n}^2)]}$$

$$\beta = \sqrt{0.5[|\tilde{n}|^2 - \text{Im}(\tilde{n}^2)]}$$

We used the eikonal approximation (assuming light propagation along straight lines)³⁹ to calculate the phase and the amplitude of the XUV field following propagation through a gas jet with thickness L :

$$A(x, z = 0) = A_0(x) e^{i\omega L[(n-1) - \beta]/c}$$

Here A_0 is the amplitude of the incoming XUV field. We restricted our calculations of the complex refraction to the xz plane, taking into account only the x dependence of the XUV beam profile. As the deflection angles observed in the present experiments are small (for example, the maximum deflection angle observed in Fig. 2a is about 3.5 mrad), we used the Kirchhoff diffraction formula in the small-angle approximation³⁹ to calculate the XUV amplitude at a screen located at a distance S from the jet. The angle-dependent XUV amplitudes (arbitrary units) were calculated with the following formula:

$$\tilde{A}(\theta, z = S) \propto \int A(x) e^{i\omega[x \sin(\theta) + x^2/(2S)]/c} dx$$

One can also estimate the focal length f of a biconvex lens with equal radii using the following formula:

$$f = \frac{R}{2(n_2 - n_1)}$$

Here R is the radius of the curved surfaces, n_2 is the refractive index of the lens material and n_1 is the refractive index of the surrounding material (that is, $n_1 = 1$ in our case). Assuming a biconvex He lens with a constant density and a radius of $R = 1$ mm, which is equivalent to the density profile used in the calculation, we estimate the focal length of the He lens. The refractive index n_2 at 20.22 eV at a gas density of 1.07×10^{20} atoms cm⁻³ (where the spot size is smallest; see Fig. 4d) is 1.000542. Using these values, we obtain a focal length of 92 cm, which is in good agreement with the results shown in Fig. 4d.

The temporal envelope of the XUV pulse was calculated by taking the square of the Fourier transform along the spectral axis in Fig. 4 and averaging over the spatial coordinate. As mentioned in the main text, the incoming pulses are assumed to be chirped by about -8 meV fs⁻¹ resulting in an XUV pulse with a duration of 24 fs. Depending on the applied gas density, the pulses in the focus are either compressed or acquire a positive chirp.

Data availability

The data that support the findings of this study are available from the corresponding authors upon request.

34. Neidel, C. et al. Probing time-dependent molecular dipoles on the attosecond time scale. *Phys. Rev. Lett.* **111**, 033001 (2013).
35. Drescher, L. et al. Communication: XUV transient absorption spectroscopy of iodomethane and iodobenzene photodissociation. *J. Chem. Phys.* **145**, 011101 (2016).
36. Galbraith, M. C. E. et al. Few-femtosecond passage of conical intersections in the benzene cation. *Nat. Commun.* **8**, 1018 (2017).
37. He, X. et al. Spatial and spectral properties of the high-order harmonic emission in argon for seeding applications. *Phys. Rev. A* **79**, 063829 (2009).
38. Gademann, G., Ple, F., Paul, P.-M. & Vrakking, M. J. J. Carrier-envelope phase stabilization of a terawatt level chirped pulse amplifier for generation of intense isolated attosecond pulses. *Opt. Express* **19**, 24922 (2011).
39. Born, M. & Wolf, E. *Principles of Optics* 7th expanded edn (Cambridge Univ. Press, Cambridge, 1999).
40. Wiese, W. L., Smith, M. W. & Glennon, B. M. *Atomic Transition Probabilities: Hydrogen through Neon*. Technical report, National Standard Reference Data System. (NBS, 1966).
41. Wiese, W. L., Smith, M. W. & Miles, B. M. *Atomic Transition Probabilities: Sodium through Calcium*. Technical report, National Standard Reference Data System (NBS, 1969).

Transformation between meron and skyrmion topological spin textures in a chiral magnet

X. Z. Yu^{1*}, W. Koshibae¹, Y. Tokunaga², K. Shibata¹, Y. Taguchi¹, N. Nagaosa^{1,3} & Y. Tokura^{1,3}

Crystal lattices with tetragonal or hexagonal structure often exhibit structural transitions in response to external stimuli¹. Similar behaviour is anticipated for the lattice forms of topological spin textures, such as lattices composed of merons and antimerons or skyrmions and antiskyrmions (types of vortex related to the distribution of electron spins in a magnetic field), but has yet to be verified experimentally^{2,3}. Here we report real-space observations of spin textures in a thin plate of the chiral-lattice magnet $\text{Co}_8\text{Zn}_9\text{Mn}_3$, which exhibits in-plane magnetic anisotropy. The observations demonstrate the emergence of a two-dimensional square lattice of merons and antimerons from a helical state, and its transformation into a hexagonal lattice of skyrmions in the presence of a magnetic field at room temperature. Sequential observations with decreasing temperature reveal that the topologically protected skyrmions remain robust to changes in temperature, whereas the square lattice of merons and antimerons relaxes to non-topological in-plane spin helices, highlighting the different topological stabilities of merons, antimerons and skyrmions. Our results demonstrate the rich variety of topological spin textures and their lattice forms, and should stimulate further investigation of emergent electromagnetic properties.

Periodic atomic arrays form crystals that can have body-centred tetragonal structures or hexagonal close-packed structures¹. These structural types are also observed in topological matter, such as in the square lattice of magnetic flux in superconductors⁴ and in the hexagonal lattice of magnetic skyrmions in chiral-lattice magnets^{5–7}. (Skyrmions are topological spin textures with topological number

$$N = \frac{1}{4\pi} \int \mathbf{n} \cdot \left(\frac{\partial \mathbf{n}}{\partial x} \times \frac{\partial \mathbf{n}}{\partial y} \right) dx dy$$

where $\mathbf{n} = \mathbf{M}/|\mathbf{M}|$ and \mathbf{M} is the magnetic moment.) The structural transition between square and hexagonal lattices has been theoretically predicted² to occur in two-dimensional chiral-lattice magnets with tuning of the material parameters, such as magnetic anisotropy. Controlling the lattice form of topological objects is important to be able to use their collective dynamics in real materials³.

We study nanometre-scale topological spin textures, such as merons (vortex-like spin textures with $N = -1/2$), antimerons and skyrmions, to confirm experimentally the transformation between the square and hexagonal lattices. Skyrmions with $N = -1$ have been discovered in chiral and polar magnets and at ferromagnetic interfaces, in which the antisymmetric spin exchange interaction arises from relativistic spin–orbit coupling (Dzyaloshinskii–Moriya interaction)^{2,3,5–15}. In a skyrmion, whole magnetic moments swirl from the south pole at the core to the north pole at the perimeter, encompassing a sphere. The hexagonal skyrmion lattice is thermodynamically stabilized by the hybridization of three helices lying within the plane perpendicular to the magnetic field in chiral-lattice magnets^{5–7}.

Merons ($N = -1/2$) and antimerons ($N = +1/2$) are topologically distinct from skyrmions ($N = -1$) and antiskyrmions¹⁶ ($N = +1$);

in merons and antimerons, the magnetic moments in the core region point upwards or downwards, and those near the perimeter align in-plane, producing degrees of freedom for magnetic helicity and polarity⁷ (Fig. 1a–d). The topological number N of skyrmions and antiskyrmions is determined by the product of the vorticity (the rotation direction of the in-plane component of the magnetic moments) and the direction of the magnetic moment in the core⁷. We classify merons and antimerons according to the sign of N .

Merons, antimerons and skyrmions, and the structural transitions of their lattices, are expected to be realized under the condition of increased magnetic anisotropy, which is also crucial for controlling topological spin textures in chiral-lattice systems². Several theoretical studies have asserted that in-plane magnetic anisotropy energetically stabilizes the square lattice of merons and antimerons, rather than the hexagonal skyrmion lattice that otherwise forms owing to the competition between the Dzyaloshinskii–Moriya and ferromagnetic interactions^{2,3,17,18}. However, merons and antimerons, and their lattice forms, have not yet been observed experimentally, although micrometre-scale artificial merons with random vorticities have been reported in permalloys^{19–22}.

Here, we report the experimental observation of merons and antimerons, and the transformation from the square (anti)meron lattice to the hexagonal skyrmion lattice. We present direct real-space imaging of topological spin textures in a thin plate of the chiral-lattice magnet $\text{Co}_8\text{Zn}_9\text{Mn}_3$, which exhibits the Dzyaloshinskii–Moriya interaction and has a high (above room temperature) transition temperature to the helical state (T_C)⁸. We also reveal the different topological stability for merons and skyrmions.

In Fig. 1e we show a schematic of the square (anti)meron lattice, which is described by a ‘double \mathbf{q} ’ structure with orthogonal \mathbf{q} vectors, where \mathbf{q} is the wavevector of in-plane helices. In Fig. 1f we present the magnetization textures observed in a (001) thin plate of $\text{Co}_8\text{Zn}_9\text{Mn}_3$ at 295 K by means of Lorentz transmission electron microscopy (TEM) with a normal magnetic field of 20 mT after a 10-mT field cooling. The white arrows and colours in Fig. 1f denote the magnitude and the direction of the in-plane magnetization, whereas the dark colour indicates the out-of-plane magnetizations²³. A comparison between Fig. 1e and Fig. 1f indicates that the experimentally observed spin texture coincides with the square (anti)meron lattice. The square (anti)meron lattice consists of periodic arrays of alternating convergence and divergence of magnetization, as theoretically predicted. Figure 1f clearly demonstrates these periodic arrays. Figure 1f also indicates that the maximum in-plane magnetization is observed at the periphery of merons and antimerons, whereas the minimum (almost zero) is at the core. In comparison with the theoretical prediction of square (anti)meron lattices (Fig. 1e), the out-of-plane magnetization at the cores of adjacent merons and antimerons with a vorticity of $+1$ are plausibly antiparallel, although the Lorentz TEM is unable to confirm the direction of the out-of-plane magnetization parallel to the incident electron beam. On the other hand, the merons with a vorticity of -1 prefer the upwards-pointing core magnetizations parallel to the external field

¹RIKEN Center for Emergent Matter Science (CEMS), Wako, Japan. ²Department of Advanced Materials Science, University of Tokyo, Kashiwa, Japan. ³Department of Applied Physics, University of Tokyo, Tokyo, Japan. *e-mail: yu_x@riken.jp

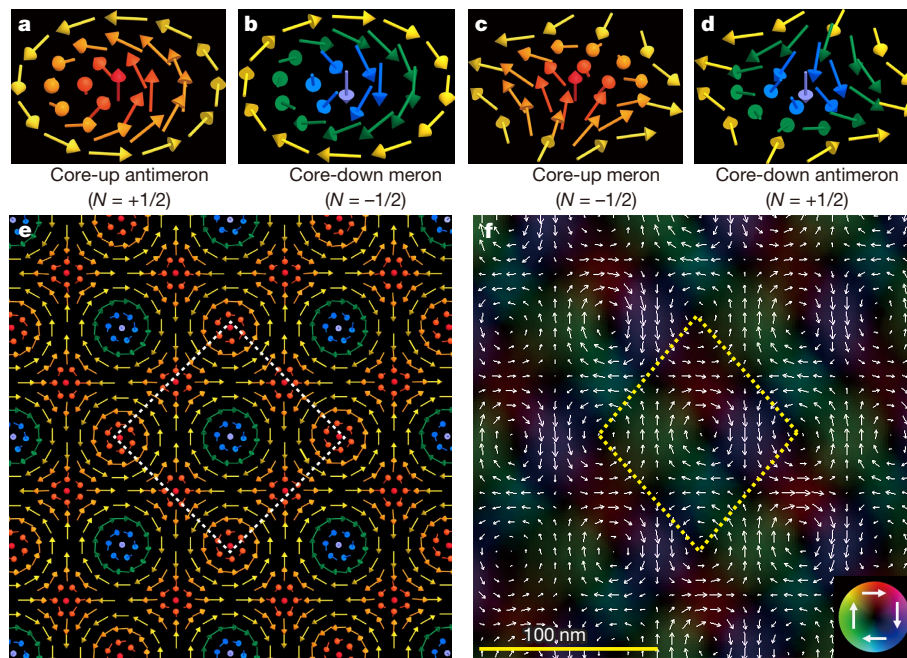


Fig. 1 | Real-space observations of a square lattice of merons and antimerons in a thin plate of the chiral-lattice magnet $\text{Co}_8\text{Zn}_9\text{Mn}_3$. **a–d**, Schematics of the magnetization textures of a core-up antimeron (**a**) and a core-down meron (**b**), each with a vorticity of $+1$, and a core-up meron (**c**) and a core-down antimeron (**d**), each with a vorticity of -1 . Coloured arrows indicate directions of the magnetic moments. **e**, Theoretically predicted square (anti)meron lattice. **f**, Real-space magnetization textures of the square (anti)meron lattice, observed by

direction, resulting in an energetically stabilized square (anti)meron lattice in the thin plate. An in situ Lorentz TEM video (Supplementary Video 1) demonstrates the formation process of the square (anti)meron lattice. Composed of merons and antimerons, it can be viewed as the hybridization of two in-plane helices; this is in contrast to the formation of the hexagonal skyrmion lattice with the hybridization of three helices under a higher field, in which no antiskyrmions appears. In comparison to the ideal case (Fig. 1e), the square (anti)meron lattice observed in a (001) $\text{Co}_8\text{Zn}_9\text{Mn}_3$ thin plate seems to elongate in the vertical direction, possibly because of the effect of strain on the spin textures in the thin plate²⁴.

We obtained experimental evidence for the transformation of the square (anti)meron lattice into the hexagonal skyrmion lattice at 295 K by finely varying the normal field. In Fig. 2a, b we show magnified snapshots from Supplementary Video 1 under 20 mT and 50 mT. The image for a higher magnetic field (60 mT; Fig. 2c) was taken separately, as a still image (see details in Methods and Extended Data Fig. 1c for the experimental procedure). Upon increasing the bias field, the real-space images and the related fast Fourier transforms (Fig. 2a–c, insets) reveal a transformation from a square lattice (Fig. 2a) to a hexagonal lattice (Fig. 2c) via a deformed square lattice (Fig. 2b) of the spin textures. The continuous deformation of the spin textures appears concurrently in the lattice transformation process.

In Fig. 2d–f we show enlarged magnetization textures in the boxed areas of Fig. 2a–c. In Fig. 2d, alternating core-down merons and core-up antimerons, accompanied by core-up merons, appear to form a square (anti)meron lattice. When the field is increased to 50 mT (Fig. 2e), the in-plane components (along the dashed lines) become small, indicating that the magnetization tends to point up. This means that the core-up merons turn into ferromagnetic regions between skyrmions, which form the deformed square lattice. Further increasing the field to 60 mT, the core-up antimerons change completely to out-of-plane field-magnetized domains, and the hexagonal skyrmion lattice is stabilized (Fig. 2f). The observed magnetization textures for square (anti)meron and hexagonal skyrmion lattices agree well with micromagnetic

using Lorentz TEM under a 20-mT field applied normally to a (001) thin plate of $\text{Co}_8\text{Zn}_9\text{Mn}_3$ after 10-mT field cooling from 340 K (above the helical-ordering transition temperature $T_C \approx 325 \text{ K}$ ³⁰). The thickness of the thin plate is approximately 100 nm; the width and length are approximately 10 μm and 50 μm , respectively. The arrows and colour scale indicate the direction and magnitude of the in-plane magnetization; black indicates out-of-plane magnetization. The dashed lines in **e** and **f** define a unit cell of the square (anti)meron lattice.

simulations (Extended Data Fig. 2). As an important feature of the transition between the lattices, the total topological number does not change during the continuous deformation of the spin texture.

In Fig. 2g–i we show a schematic of the transformation that we observe. In the square (anti)meron lattice (Fig. 2g), the total topological number N_{total} in the unit (bounded by dashed lines) composed of one core-down meron ($N = -1/2$), four quarters of core-up antimerons ($N = +1/2$) and four halves of core-up merons ($N = -1/2$) is $N_{\text{total}} = -1/2 + (+1/2) \times (1/4) \times 4 + (-1/2) \times (1/2) \times 4 = -1$. As the bias field increases, the square (anti)meron lattice deforms continuously to the square skyrmion lattice, accompanied by the transformation of core-up merons and antimerons to field-magnetized moments surrounding a core-down skyrmion (Fig. 2h). The topological number for the unit (bounded by dashed lines) composed of one skyrmion is -1 . Further increasing the bias field possibly induces a transformation from the square to the hexagonal skyrmion lattice, while keeping the topological number of -1 in the unit (bounded by dashed lines; Fig. 2i). This interpretation is plausible for the transformation observed here, indicating the topological invariance of spin textures.

To examine the stability of the meron and the antimeron in the thin plate, we created a square (anti)meron lattice at 295 K following the procedure described above and then cooled the thin plate to temperatures far below T_C while keeping the bias field of 20 mT. In Fig. 3a–d we show the Lorentz TEM images observed at several temperatures in this procedure. The same experimental protocol was used to check the stability of the hexagonal skyrmion lattice field-cooled at 60 mT. In Fig. 3e–h we show the skyrmion lattice observed at various temperatures during the field cooling. At 295 K, we observe the square (anti)meron lattice at 20 mT (Fig. 3a) and the hexagonal skyrmion lattice at 60 mT (Fig. 3e). With the decrease in temperature, the square (anti)meron lattice becomes deformed and finally collapses to almost in-plane helices at about 120 K; at this temperature, the non-topological helical and conical phases are dominant as the thermodynamically stable states in the thin plate under zero and 10-mT field cooling (Extended Data Fig. 1). In contrast to the destabilization of the square (anti)meron lattice at lower

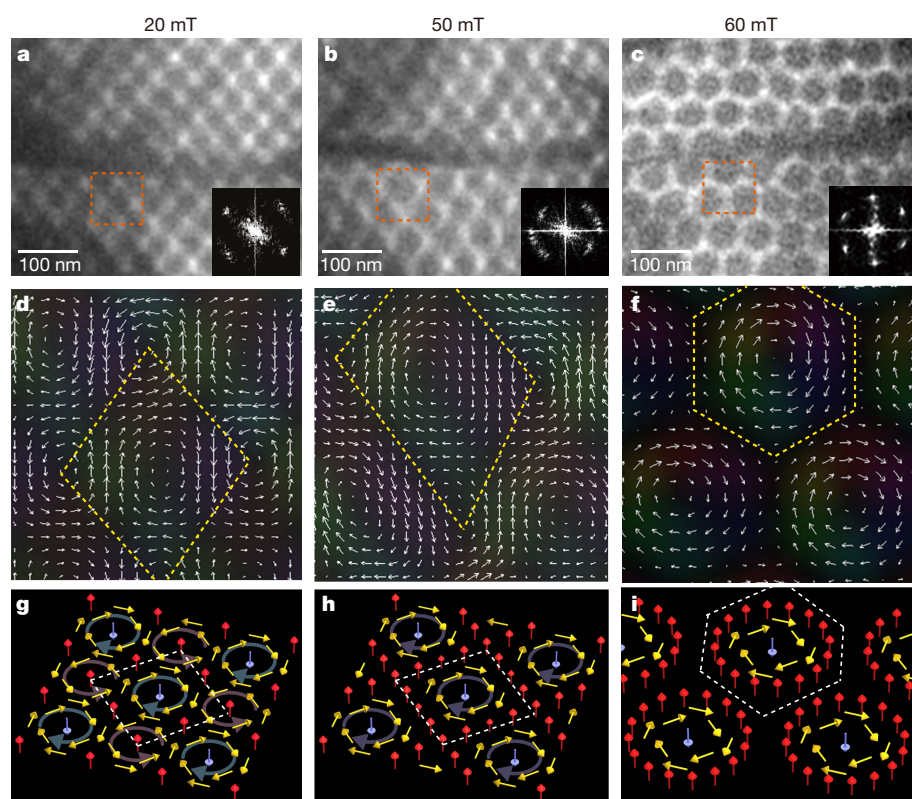


Fig. 2 | Magnetically induced transformation of a square (anti)meron lattice to a hexagonal skyrmion lattice via a deformed skyrmion lattice in the (001) plate of $\text{Co}_8\text{Zn}_9\text{Mn}_3$ at 295 K. **a–c**, Over-focused Lorentz TEM images and related fast Fourier transforms (insets) observed with increasing field, while keeping the temperature at 295 K after 10-mT field cooling (the experimental procedure is denoted by a black arrow in the phase diagram in Extended Data Fig. 1c). The Lorentz TEM image in **c** was obtained by tilting the thin plate slightly (about 2°) to remove the diffractive contrast and hence increase the intensity of the magnetic signal. The bias field affects the condition of the electron beam slightly because the field is below 100 mT for our observations. **d–f**, Magnetization textures obtained by analysing the boxed regions of the Lorentz TEM images in **a–c**, respectively, using the software Qpt²³, which is based on the transport-of-intensity equation (see Methods). Colours and arrows as in Fig. 1f. **g–i**, Schematics of the magnetization textures for the square (anti)meron lattice (**g**), the deformed square skyrmion lattice (**h**) and the hexagonal skyrmion lattice (**i**). The dashed regions in **d–i** indicate the units of the spin textures, each with a topological number of $N = -1$.

temperatures, the hexagonal skyrmion lattice survives, and no topological phase transition occurs with decreasing temperature, although the shape of the individual skyrmion becomes deformed, possibly owing to the effect of anisotropic strain. The magnetization textures of the boxed regions in Fig. 3d (20 mT) and Fig. 3h (60 mT) at 120 K are shown in

Fig. 3i and Fig. 3j, respectively, with the colour wheel indicating the in-plane magnetization direction. These images confirm that only a few core-up antimerons remain in the matrix of in-plane helices with orthogonal \mathbf{q} vectors at 20 mT and 120 K. By contrast, in the case of field cooling at 60 mT, the hexagonal deformed-skyrmion lattice exists

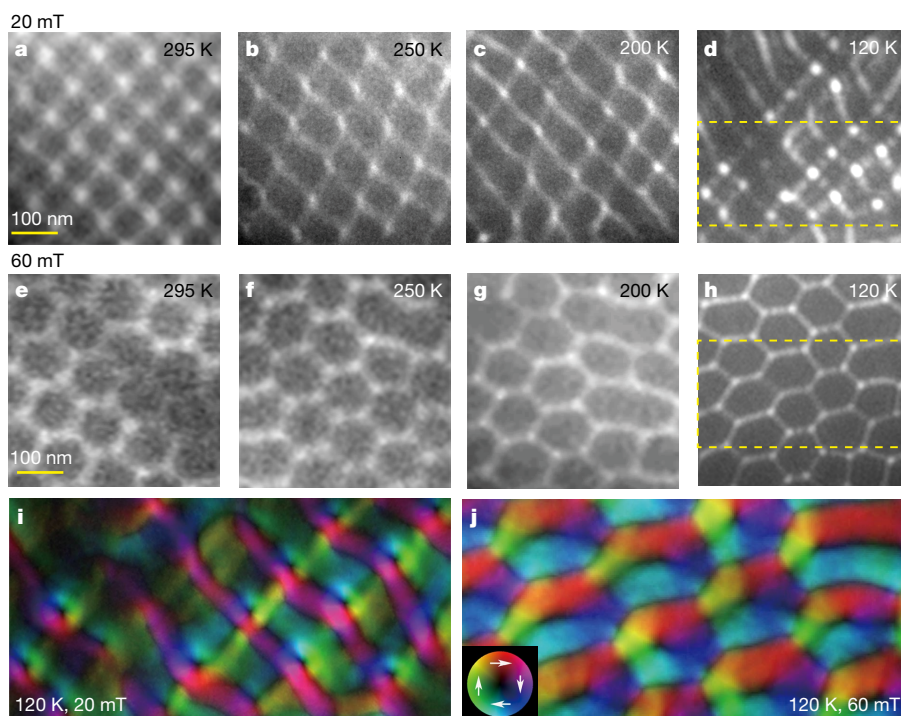


Fig. 3 | Stability of the square (anti)meron and hexagonal skyrmion lattices in the (001) plate of $\text{Co}_8\text{Zn}_9\text{Mn}_3$. **a–h**, Over-focused Lorentz TEM images of the square (anti)meron (**a–d**) and hexagonal skyrmion (**e–h**) lattices observed with decreasing temperature at a magnetic field of 20 mT and 60 mT, respectively (experimental procedures are denoted by

red dashed arrows in the phase diagram in Extended Data Fig. 3b). **i, j**, Magnetization textures for the boxed regions in **d** and **h**, respectively, deduced by analysing Lorentz TEM images with the transport-of-intensity equation. The colour wheel in the inset of **j** indicates the in-plane magnetization direction.

robustly at the same temperature (see also the temperature–field phase diagram in Extended Data Fig. 3). The result for the field cooling at 60 mT is in accord with previous experimental studies on topologically protected skyrmions in chiral-lattice magnets^{25–29}.

With in-plane anisotropy (Methods, Extended Data Figs. 1, 5, 6), it is natural to anticipate the formation of merons and antimerons because of the consistent vorticities between the neighbouring spin textures³. In addition, the region surrounded by the two merons and two antimerons has a vorticity of -1 and the upwards-pointing magnetic field generates core-up moments, which results in the meron (Fig. 2g). This is why the square (anti)meron lattice forms at weak magnetic fields. Although metastability is also discerned for the square (anti)meron lattice beyond thermal equilibrium (compare Extended Data Figs. 1c and 3b), the (anti)meron lattice is easier to deform to a helix than is the skyrmion lattice, because the total magnetization and each topological number are smaller, so the potential barrier to change the topological number is expected to be smaller. In this sense, we consider the square (anti)meron lattice to be located between a helix and a skyrmion lattice.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0745-3>.

Received: 12 July 2018; Accepted: 11 October 2018;

Published online 5 December 2018.

- Kittel, C. *Introduction to Solid State Physics* Ch. 1 (John Wiley & Sons, New York, (2005).
- Lin, S. Z., Saxena, A. & Batista, C. D. Skyrmion fractionalization and merons in chiral magnets with easy-plane anisotropy. *Phys. Rev. B* **91**, 224407 (2015).
- Yi, S. D., Onoda, S., Nagaosa, N. & Han, J. H. Skyrmions and anomalous Hall effect in a Dzyaloshinskii–Moriya spiral magnet. *Phys. Rev. B* **80**, 054416 (2009).
- Berdiyev, G. R., Milošević, M. V. & Peeters, F. M. Vortex configurations and critical parameters in superconducting thin films containing antidote arrays: nonlinear Ginzburg–Landau theory. *Phys. Rev. B* **74**, 174512 (2006).
- Mühlbauer, S. et al. Skyrmion lattice in a chiral magnet. *Science* **323**, 915–919 (2009).
- Yu, X. Z. et al. Real-space observation of a two-dimensional skyrmion crystal. *Nature* **465**, 901–904 (2010).
- Nagaosa, N. & Tokura, Y. Topological properties and dynamics of magnetic skyrmions. *Nat. Nanotechnol.* **8**, 899–911 (2013).
- Tokunaga, Y. et al. A new class chiral materials hosting magnetic skyrmions beyond room temperature. *Nat. Commun.* **6**, 7638 (2015).
- Heinze, S. et al. Spontaneous atomic-scale magnetic skyrmion lattice in two dimensions. *Nat. Phys.* **7**, 713–718 (2011).
- Kézsmárki, I. et al. Néel-type skyrmion lattice with confined orientation in the polar magnetic semiconductor GaV₄S₈. *Nat. Mater.* **14**, 1116–1122 (2015).
- Li, W. et al. Emergence of skyrmions from rich parent phases in the molybdenum nitrides. *Phys. Rev. B* **93**, 060409(R) (2016).
- Woo, S. et al. Observation of room-temperature magnetic skyrmions and their current-driven dynamics in ultrathin metallic ferromagnets. *Nat. Mater.* **15**, 501–506 (2016).
- Jiang, W. et al. Blowing magnetic skyrmion bubbles. *Science* **349**, 283–286 (2015).
- Sampaio, J., Cros, V., Rohart, S., Thiaville, A. & Fert, A. Nucleation, stability and current-induced motion of isolated magnetic skyrmions in nanostructures. *Nat. Nanotechnol.* **8**, 839–844 (2013).
- Zheng, F. S. et al. Experimental observation of chiral magnetic bobbins in B20-type FeGe. *Nat. Nanotechnol.* **13**, 451–455 (2018).
- Nayak, A. K. et al. Magnetic antiskyrmions above room temperature in tetragonal Heusler materials. *Nature* **548**, 561–566 (2017).
- Ozawa, R. et al. Vortex crystals with chiral stripes in itinerant magnets. *J. Phys. Soc. Jpn* **85**, 103703 (2016).
- Vousden, M. et al. Skyrmions in thin films with easy-plane magnetocrystalline anisotropy. *Appl. Phys. Lett.* **108**, 132406 (2016).
- Shinjo, T., Okuno, T., Hassdorf, R., Shigeko, K. & Ono, T. Magnetic vortex core observation in circular dots of permalloy. *Science* **289**, 930–932 (2000).
- Phatak, C., Petford-Long, A. K. & Heinonen, O. Direct observation of unconventional topological spin structure in coupled magnetic discs. *Phys. Rev. Lett.* **108**, 067205 (2012).
- Wintz, S. et al. Topology and origin of effective spin meron pairs in ferromagnetic multilayer elements. *Phys. Rev. Lett.* **110**, 177201 (2013).
- Tan, A. et al. Topology of spin meron pairs in coupled Ni/Fe/Co/Cu (001) disks. *Phys. Rev. B* **94**, 014433 (2016).
- Ishizuka, K. & Allman, B. Phase measurement in electron microscopy using the transport of intensity equation. *J. Electron Microsc.* **54**, 191–197 (2005).
- Shibata, K. et al. Large anisotropic deformation of skyrmions in strained crystals. *Nat. Nanotechnol.* **10**, 589–592 (2015).
- Karube, K. et al. Robust metastable skyrmions and their triangular–square lattice structural transition in a high-temperature chiral magnet. *Nat. Mater.* **15**, 1237–1242 (2016).
- Nakajima, T. et al. Skyrmion lattice structural transition in MnSi. *Sci. Adv.* **3**, e1602562 (2017).
- Oike, H. et al. Interplay between topological and thermodynamic stability in a metastable magnetic skyrmion lattice. *Nat. Phys.* **12**, 62–66 (2016).
- Münzer, W. et al. Skyrmion lattice in the doped semiconductor Fe_{1-x}Co_xSi. *Phys. Rev. B* **81**, 041203(R) (2010).
- Yu, X. Z. et al. Aggregation and collapse dynamics of skyrmions in a non-equilibrium state. *Nat. Phys.* **14**, 832–836 (2018).
- Yu, X. Z. et al. Current-induced nucleation and annihilation of magnetic skyrmions at room temperature in a chiral magnet. *Adv. Mater.* **29**, 1606178 (2017).

Acknowledgements We thank M. Ishida, Á. Butykai, D. Morikawa, T.-H. Arima and M. V. Mostovoy for experimental support and discussions. N.N. was supported by JSPS KAKENHI (grant numbers JP26103006 and JP18H03676) and JST CREST (grant number JPMJCR1874), Japan.

Reviewer information Nature thanks S. Woo and the other anonymous reviewers for their contribution to the peer review of this work.

Author contributions Y. Tokura conceived the project. X.Z.Y. performed Lorentz TEM and analysed the experimental data. W.K. and N.N. performed the theoretical analyses. Y. Tokunaga and Y. Taguchi synthesized the Co–Zn–Mn alloys. K.S. simulated the Lorentz TEM images. All authors discussed the data and collaborated on the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0745-3>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0745-3>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to X.Z.Y.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Sample preparation and experimental methods. Bulk Co-Zn-Mn alloys were synthesized using procedures described elsewhere^{8,30}. The thin plates for the Lorentz TEM observations were prepared by an Ar-ion milling process after mechanical polishing of the bulk sample. Real-space observations were performed using a multifunctional TEM (JEM2800, JEOL) equipped with single-tilt heating/cooling (HC 3500) and double-tilt nitrogen cooling (Gatan 636) holders. The temperature of the thin plate was controlled between 95 K and 350 K. The magnetic configurations were obtained by Lorentz TEM observations.

Characterization of magnetization textures in a thin plate of Co₈Zn₉Mn₃. Lorentz TEM is useful for observing the in-plane magnetic configurations in thin magnets. The convergence (divergence) of the electron beam reflecting the in-plane magnetization distribution can be imaged in the defocused Lorentz TEM images as bright (dark) contrast because of the interactions between the incident electrons and the in-plane magnetic components in thin magnets. Such contrasts should reverse in under-focused and over-focused Lorentz TEM images. Therefore, an ideal screw structure that follows the direction of the in-plane wavevector can be projected as alternating bright and dark stripes with a constant period. For skyrmions that exhibit clockwise (anticlockwise) helicity, the bright (dark) dots should appear in the defocused image plane. We extracted electron-phase images and then converted them to magnetization textures to avoid blurring and artificial background noise caused by the surface roughness of the sample using the software Qpt²³, which is based on the transport-of-intensity equation:

$$\frac{2\pi}{\lambda} \frac{\partial I(x, y, z)}{\partial z} = -\nabla_{xy} \cdot [I(x, y, z) \nabla_{xy} \varphi(x, y, z)] \quad (1)$$

where λ , $I(x, y, z)$ and $\varphi(x, y, z)$ represent the wavelength, intensity and phase of the incident electron beam, respectively. The gradient of the electron-beam intensity ($\partial I(x, y, z)/\partial z$) in the under-focused and over-focused Lorentz TEM images enables us to obtain $\nabla_{xy} \varphi(x, y, z)$. The Maxwell–Ampère equation establishes a relationship between $\varphi(x, y, z)$ and the magnetization M :

$$\nabla_{xy} \varphi(x, y, z) = -\frac{e}{h} (M \times n) t \quad (2)$$

where t and n are the sample thickness and unit vector perpendicular to the sample surface, respectively.

Crystal structure, magnetic structure and magnetic phase diagrams observed in a thin plate of Co₈Zn₉Mn₃ with a low-field (10 mT) cooling procedure. The Co₈Zn₉Mn₃ alloy crystallizes in a chiral cubic structure with space group $P4_132$ or $P4_332$ (Extended Data Fig. 1a, b)⁸. Extended Data Fig. 1c shows the magnetic phase diagram of the thin Co₈Zn₉Mn₃ obtained from field-increasing runs after low-field (<10 mT) cooling. The hexagonal skyrmion lattice (hex-SkL) phase (red-coloured region in Extended Data Fig. 1c) and the square lattice of merons and antimerons (sq-ML) (green-coloured region in Extended Data Fig. 1c) are thermodynamically stabilized in a narrow window near T_C . The helical structure at zero field and the conical structure under a finite bias field are dominant at lower temperatures. Extended Data Fig. 1d demonstrates the striped domains of in-plane helices at a lower temperature (for example, 95 K); when the temperature is increased to 295 K (Extended Data Fig. 1e), the striped domains turn into multi-domains composed of helices with the in-plane wavevector along the [100] axis (alternating red and green stripes), and possible helices with the out-of-plane wavevector along the [001] axis (dark areas). Further increasing the temperature above 300 K to approach T_C (about 325 K), the in-plane helices disappear and the dark area in Extended Data Fig. 1e (possible helices with the out-of-plane wavevector along the [001] axis) expands over the view area, revealing the planar magnetic anisotropy in the thin plate of Co₈Zn₉Mn₃. With the application of a normal field of 65 mT, a hex-SkL appears at 300 K (Extended Data Fig. 1f).

Simulations of magnetic configurations. In Extended Data Fig. 2, the distribution of the unit vector of the magnetic moment $n = (n_x, n_y, n_z)$ for Fresnel (defocused) TEM image simulation was prepared by approximating the magnetic configurations with the superposition of proper-screw-type helices as follows. For the square lattice, we (1) prepare two proper-screw helices with q vectors perpendicular to each other; (2) add each n component of the proper-screw helices; (3) add a uniform positive value n_z^{add} to the n_z component; and (4) normalize the length of the vectors at each site so that $|n| = (n_x^2 + n_y^2 + n_z^2)^{1/2} = 1$. Owing to step (3), core-up merons are selected and no core-down antimerons appear, as in Fig. 1e. For the hexagonal lattice, we (1) prepare three proper-screw helices q vectors at relative angles of $2\pi/3$ to each other; (2) add each n component of the proper-screw helices; and (3) normalize the length of the vectors at each site so that $|n| = (n_x^2 + n_y^2 + n_z^2)^{1/2} = 1$.

Fresnel image (defocused Lorentz TEM image) simulation. We performed the Fresnel image (defocused Lorentz TEM) simulation for the magnetic configurations using a custom-made program. The magnetic phase shift was calculated using a Fourier approach³¹; parameters are listed in Extended Data Table 1.

We did not consider contributions to the electron phase from a stray magnetic induction field in the vacuum region. Although the parameters for the simulation, such as the dimensions of the magnetic configurations, the magnetization, the defocus length and the dynamic range of the images, are not exact and deviate from the experimental conditions, the resultant intensity distribution and magnetization maps are anticipated to be qualitatively correct.

It is hard to distinguish the sq-ML from the square skyrmion lattice using Fresnel images alone, because some different magnetic configurations can yield the same intensity distribution in Fresnel images. Therefore, the simulation alone cannot exclude the possibility of alternative magnetic configurations.

Various (meta)stable states realized by the processes in the temperature–magnetic-field plane in the thin plate of Co₈Zn₉Mn₃. Extended Data Fig. 3 shows the phase diagrams and several relevant Lorentz TEM images observed in the (001) thin plate of Co₈Zn₉Mn₃. In Extended Data Fig. 3a, the field cooling (FC) with $B = 60$ mT from above T_C to various temperatures below T_C followed by the increase in B results in the deformed hex-SkL at each temperature (red dashed lines). Similarly, after FC with $B = 60$ mT down to $T = 100$ K, then reducing (black arrow) B to 40 mT and subsequently raising the temperature (dashed blue line), we observe the deformed skyrmions robustly (Extended Data Figs 3c, d), but no sq-ML state. These results agree with previous experimental studies of metastabilized skyrmions in bulky and flaky chiral-lattice magnets^{25,26,32}.

The cases of 20-mT and 40-mT FC (Extended Data Fig. 3b) are different from that of 60-mT FC where the expansion of the hex-SkL phase was observed in the T – B plane. The 20-mT and 40-mT FC result in the appearance of the distinctive sq-ML phase (Extended Data Fig. 3e). However, as the magnitude of the cooling field is increased above 50 mT, the sq-ML is replaced by the hex-SkL. The sq-ML is also metastabilized with the 20-mT and 40-mT FC down to 150 K, but tends to collapse into the helical state with further decrease in T (for example, 100 K; Extended Data Fig. 3f). The helical state, once converted from the metastable sq-ML at lower temperature, always remains stable up to room temperature as long as B is less than 10 mT (see the low-field region along the abscissa in Extended Data Fig. 3b).

Various periodic arrays of the topological spin textures in the thin plate of Co₈Zn₉Mn₃. Various periodic arrays of topological spin textures, including the sq-ML at a lower field of 20-mT FC (Extended Data Fig. 4a–c), the hex-SkL at a field of 65-mT FC (Extended Data Fig. 4d–f) and skyrmion chains at a higher field of 180-mT FC (Extended Data Fig. 4g–i), are generated in a (001) thin plate of Co₈Zn₉Mn₃ at a temperature of 295 K. Lorentz TEM images (Extended Data Fig. 4a, d), the fast Fourier transforms (insets in Extended Data Fig. 4a, d) and related magnetization textures (Extended Data Fig. 4b, e) clearly represent the sq-ML with four-fold symmetry at the lower field and the hex-SkL with six-fold symmetry at the higher field, respectively. In addition to these lattice forms of topological spin textures, the curved skyrmion chains (Extended Data Fig. 4g, h) can be generated via a higher-field (180 mT) cooling procedure, indicating fertile topological states in the thin plate of Co₈Zn₉Mn₃. The magnified images of the topological spin textures (Extended Data Fig. 4c, f, i) reveal the differences in the topological number, shape and size between merons and skyrmions.

Changes in spontaneous magnetic structure in thin plates of Co-Zn-Mn with varying Mn composition. To characterize the in-plane magnetic anisotropy of the thin plates of Co-Zn-Mn with different T_C (T_C increases with decreases in the Mn composition), we performed a series of Lorentz TEM observations by varying the Mn compositions at zero field and 95 K, below the magnetic ordering temperature T_C . The electron-phase images (Extended Data Fig. 5i–l) were obtained by analysing defocused Lorentz TEM images (over-focused images shown in Extended Data Fig. 5a–d and under-focused images shown in Extended Data Fig. 5e–h) using the software Qpt. As described above, the in-plane magnetization can be evaluated from the phase shift according to equation (2). The white and dark contrasts with gradient in the electron-phase images are related to the in-plane magnetizations with opposite directions. In contrast to the ideal screw structure observed in Co₈Zn₈Mn₄ with a lower T_C of about 300 K, the band-shaped in-plane ferromagnetic domains (Extended Data Fig. 5l) appear in the CoZn thin plate with the higher T_C of about 460 K⁸. The decrease in Mn composition results in unbalanced domains with positive (white contrast) or negative (dark contrast) phase shift, showing that the magnetization stays longer in-plane compared with out-of-plane (compare Extended Data Fig. 5i and l). In other words, decreasing the Mn composition leads to the enhancement of in-plane magnetic anisotropy in Co-Zn-Mn thin plates^{18,33,34}.

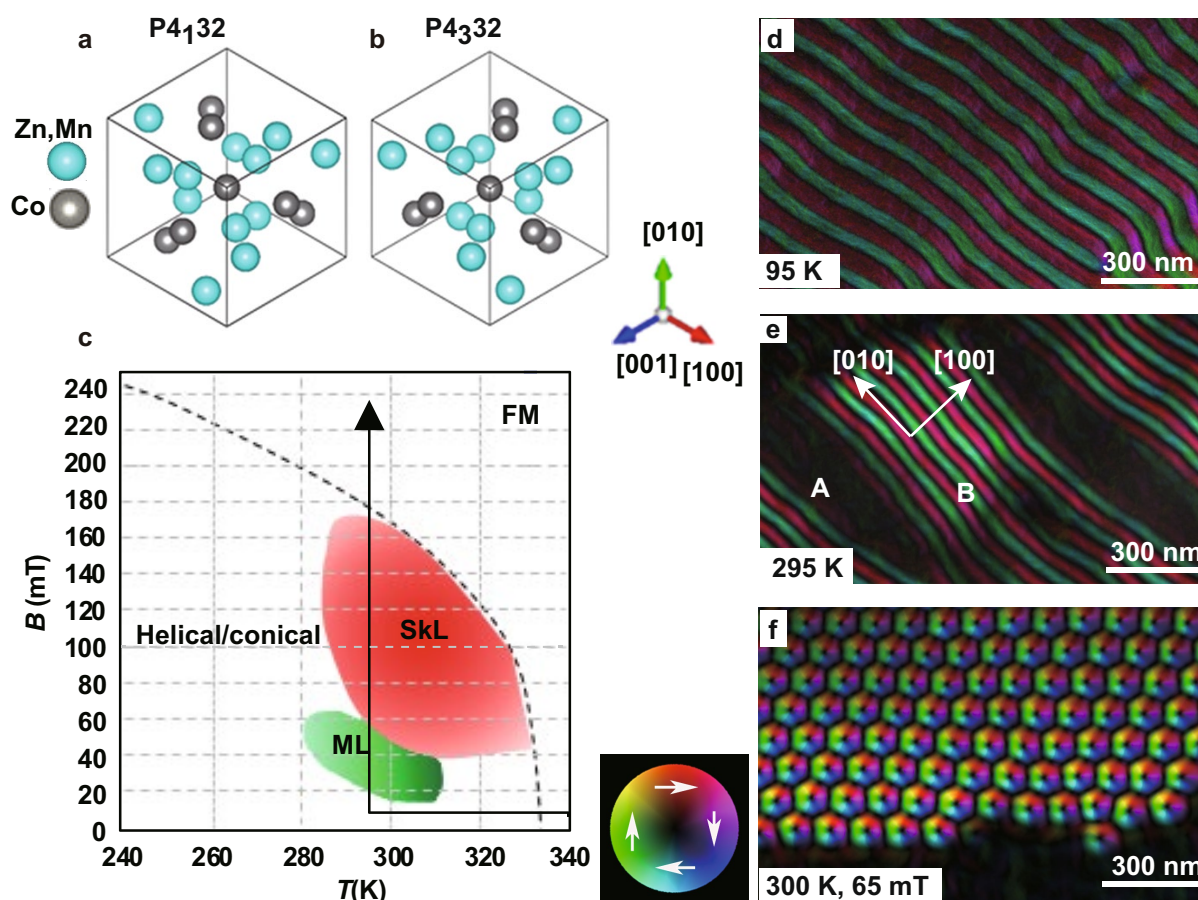
Exotic topological spin textures in thin plates of Co-Zn-Mn. We observed various exotic topological states (Extended Data Fig. 6) in the thin plates of Co-Zn-Mn with the application of magnetic fields normal to the plates. Extended Data Fig. 6a shows a Lorentz TEM image of skyrmion chains with a 90-mT field at room temperature in Co₈Zn₉Mn₃. Such unfavourable spin textures in chiral-lattice systems in terms of the Dzyaloshinskii–Moriya interaction may become plausible in thin films with in-plane magnetic anisotropy^{3,18}. With decreases in Mn composition, we observed the structure of two bound, deformed skyrmions of opposite helicity in

$\text{Co}_8\text{Zn}_{10}\text{Mn}_2$ (Extended Data Fig. 6b) and bubble-like domains in CoZn (Extended Data Fig. 6c), which points to the further enhancement of the in-plane anisotropy with a decrease in Mn composition in thin plates of Co-Zn-Mn.

Data availability

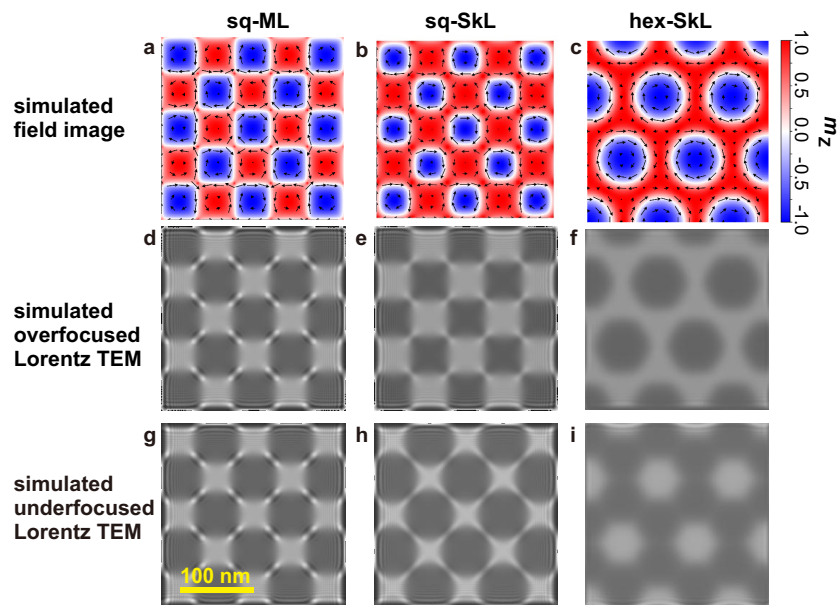
The data shown in the figures and that support the findings of this study are available from the corresponding author on reasonable request.

31. Beleggia, M. et al. Quantitative study of magnetic field distribution by electron holography and micromagnetic simulations. *Appl. Phys. Lett.* **83**, 1435 (2003).
32. Morikawa, D. et al. Deformation of topologically-protected supercooled skyrmions in a thin plate of chiral magnet $\text{Co}_8\text{Zn}_8\text{Mn}_4$. *Nano Lett.* **17**, 1637–1641 (2017).
33. Bogdanov, A. & Hubert, A. Thermodynamically stable magnetic vortex states in magnetic crystals. *J. Magn. Magn. Mater.* **138**, 255–269 (1994).
34. Hubert, A. & Schäfer, R. *Magnetic Domains* Chs. 2, 3 (Springer, Berlin, 1998).



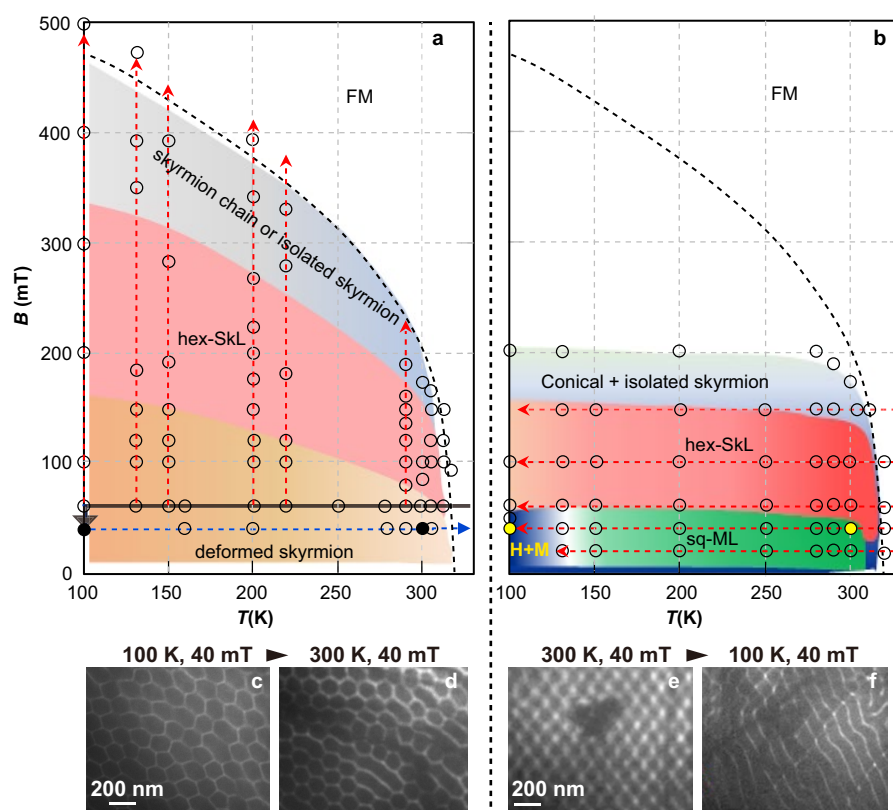
Extended Data Fig. 1 | The crystal structure, magnetic configurations and magnetic phase diagrams of the (001) thin plate of $\text{Co}_8\text{Zn}_9\text{Mn}_3$. **a, b,** Schematics of the crystal structure with space group $P4_132$ (**a**) and $P4_332$ (**b**). Coloured arrows indicate the crystal axes. **c,** Magnetic phase diagram (approximate) of the hex-SkL³⁰ and sq-ML observed over field-increasing runs from low (less than 10 mT) field cooling for a (001) thin plate of $\text{Co}_8\text{Zn}_9\text{Mn}_3$. The phase determination was based on the continuous magnetic-field scans at fixed temperatures in intervals of $\Delta T = 5$ K. The arrow indicates the field-increasing run for the Lorentz

TEM images shown in Fig. 2a–c. FM, field-magnetized ferromagnetic structure. **d, e,** Periodic stripe domains with a single wavevector along the $[100]$ axis at 95 K (**d**), the helical structure with possible multi-domains composed of helices with in-plane wavevectors (area B) and with out-of-plane wavevectors (dark regions; area A) at 295 K (**e**), respectively. **f,** A hex-SkL realized under 65 mT at 300 K. Colours in **d–f** (see colour wheel) depict the direction (white arrows) of the local in-plane magnetization; black shows the out-of-plane magnetization.



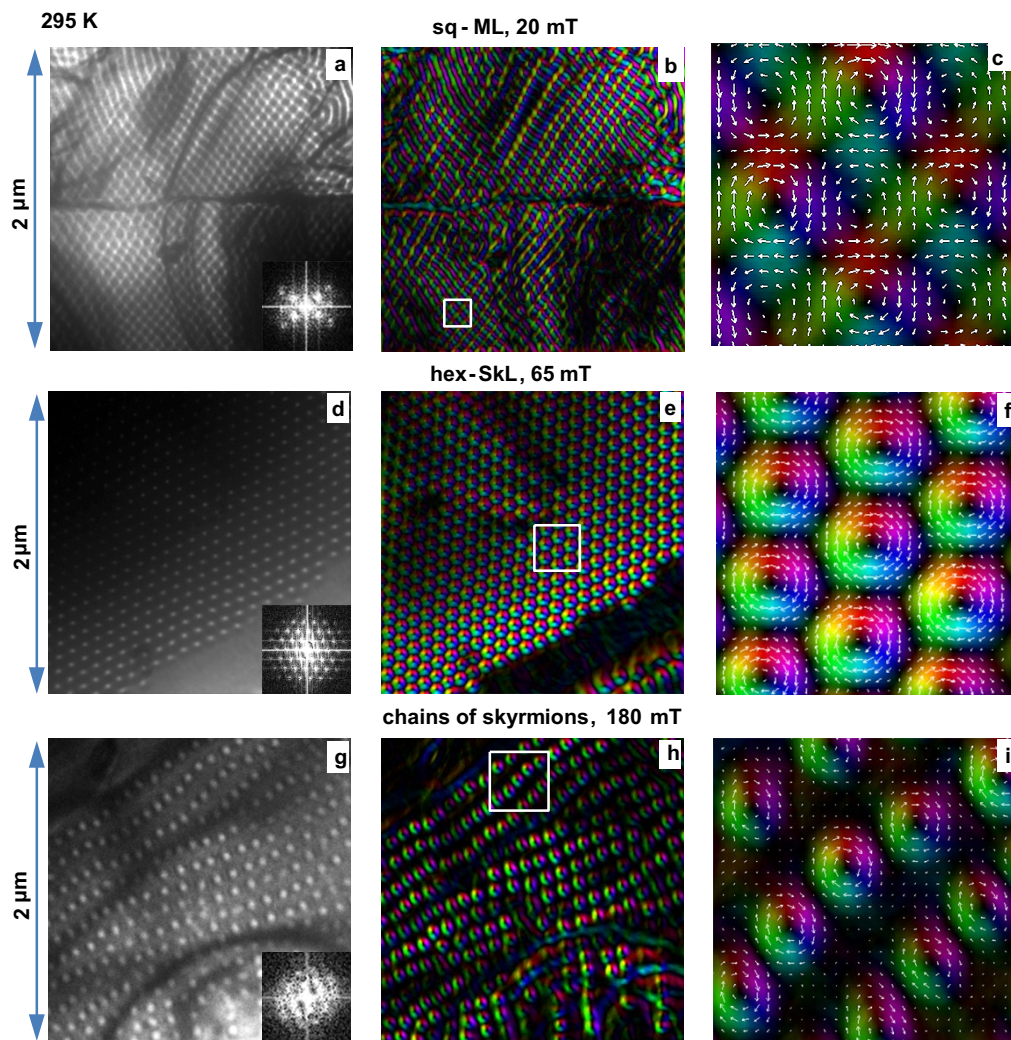
Extended Data Fig. 2 | The approximate in-plane magnetization textures and simulated defocused Lorentz TEM images. a, d, h, sq-ML. b, e, i, sq-SkL. c, f, j, hex-SkL. The parameters for the simulations are

shown in Extended Data Table 1. The colour bar indicates the normalized component of the out-of-plane magnetization m_z .



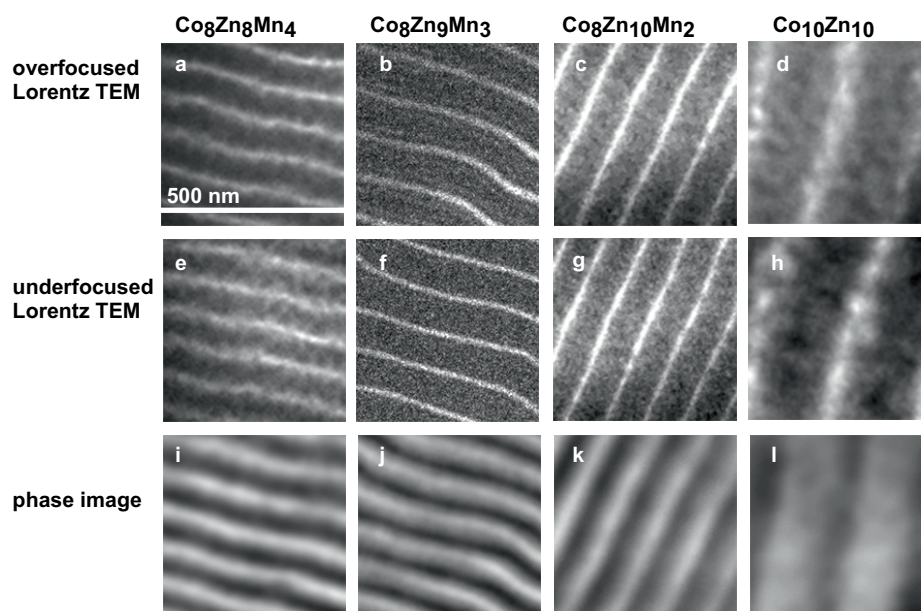
Extended Data Fig. 3 | Magnetic phase diagrams and several over-focused Lorentz TEM images observed in the (001) thin plate of $\text{Co}_5\text{Zn}_9\text{Mn}_3$ with varying temperature T and external magnetic field B . **a**, Phase diagram of the magnetic structure observed after 60-mT field cooling with increasing B (red dashed arrows), decreasing (black arrow) B and then increasing T (blue dashed arrow). **b**, Phase diagram of the magnetic structure observed after field cooling with various cooling fields

(indicated by red dashed arrows). $H + M$ shows the mixed structure of helices (dominant) and merons (minor). The open circles specify the (T, B) points that we measured. The dark blue region shows the helical phase. **c–f**, Over-focused Lorentz TEM images observed for different T and B , indicated by black solid circles in **a** (**c, d**) and yellow solid circles in **b** (**e, f**).



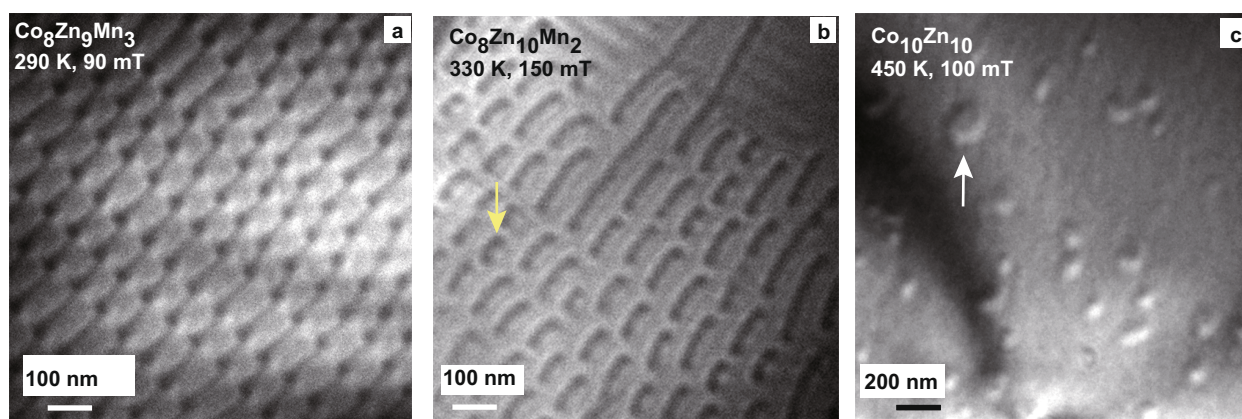
Extended Data Fig. 4 | Various periodic arrays of the topological spin textures observed in the (001) thin plate of $\text{Co}_8\text{Zn}_9\text{Mn}_3$ with varying external magnetic field. a, b, d, e, g, h, Lorentz TEM images (a, d and g; insets show the corresponding fast Fourier transforms) and their

magnetization maps (b, e and h) for the sq-ML (a, b), hex-SkL (d, e) and skyrmion chains (g, h) observed at 295 K and various fields. c, f, i, Magnified magnetization textures in the boxed areas in b, e and h.



Extended Data Fig 5 | Spontaneous magnetic structures in thin plates of Co-Zn-Mn with various Mn compositions. a–h, Defocused Lorentz TEM images observed in the thin plates of Co-Zn-Mn at zero field and 95 K.

i–l, Electron-phase images obtained from analysing Lorentz TEM images in **a–h** with the transport-of-intensity equation.



Extended Data Fig. 6 | Exotic topological spin textures in thin plates of Co-Zn-Mn with various Mn compositions. a–c, Over-focused Lorentz TEM images of skyrmion chains observed in $\text{Co}_8\text{Zn}_9\text{Mn}_3$ (a), bound

skyrmions in $\text{Co}_8\text{Zn}_{10}\text{Mn}_2$ (b; such as that indicated by the yellow arrow) and bubble-like domains in $\text{Co}_{10}\text{Zn}_{10}$ (c; such as that indicated by the white arrow).

Extended Data Table 1 | Parameters for Fresnel image simulations

Parameter	Value
M (at 300 K)	$5.7\mu_B/\text{f.u.}$
v	0.26 nm^3
B	$\mu_0 M/v = 0.26\text{ T}$
t	100 nm
V_{acc}	200 kV
C_s	1 mm
Δf	$30\text{ }\mu\text{m}$

M , μ_B , μ_0 , v , B , t , V_{acc} , C_s and Δf denote the local magnetic moment per formula unit, Bohr magneton, vacuum permeability, the unit-cell volume of the crystal lattice, the magnetic induction, the sample thickness, the accelerating voltage of the microscope, the spherical aberration of the microscope and the defocus depth of the Fresnel image in Lorentz TEM mode, respectively.

Industrial and agricultural ammonia point sources exposed

Martin Van Damme^{1,3*}, Lieven Clarisse^{1,3*}, Simon Whitburn¹, Juliette Hadji-Lazaro², Daniel Hurtmans¹, Cathy Clerbaux^{1,2} & Pierre-François Coheur¹

Through its important role in the formation of particulate matter, atmospheric ammonia affects air quality and has implications for human health and life expectancy^{1,2}. Excess ammonia in the environment also contributes to the acidification and eutrophication of ecosystems^{3–5} and to climate change⁶. Anthropogenic emissions dominate natural ones and mostly originate from agricultural, domestic and industrial activities⁷. However, the total ammonia budget and the attribution of emissions to specific sources remain highly uncertain across different spatial scales^{7–9}. Here we identify, categorize and quantify the world's ammonia emission hotspots using a high-resolution map of atmospheric ammonia obtained from almost a decade of daily IASI satellite observations. We report 248 hotspots with diameters smaller than 50 kilometres, which we associate with either a single point source or a cluster of agricultural and industrial point sources—with the exception of one hotspot, which can be traced back to a natural source. The state-of-the-art EDGAR emission inventory¹⁰ mostly agrees with satellite-derived emission fluxes within a factor of three for larger regions. However, it does not adequately represent the majority of point sources that we identified and underestimates the emissions of two-thirds of them by at least one order of magnitude. Industrial emitters in particular are often found to be displaced or missing. Our results suggest that it is necessary to completely revisit the emission inventories of anthropogenic ammonia sources and to account for the rapid evolution of such sources over time. This will lead to better health and environmental impact assessments of atmospheric ammonia and the implementation of suitable nitrogen management strategies.

Considerable effort goes into establishing spatially and temporally resolved ammonia (NH₃) bottom-up emission inventories, as these are critical drivers of models that are used to assess NH₃ distributions and impacts on the environment. Bottom-up inventories are built from activity data coupled with estimated emission factors. The correctness of these input data is their Achilles' heel, as activity data can be absent or outdated and estimated emission factors are based on specific case studies and may not be representative of either local or global conditions^{11,12}. When they are available, global measured atmospheric distributions of trace gas concentrations allow us to retrieve source emissions. In the past few years, satellite sounders have offered (bi) daily global NH₃ measurements^{13–17}, which have a huge potential to improve our knowledge of the NH₃ emission budget¹⁸. The first global distributions¹³ and inverse modelling efforts¹⁹ have confirmed the correctness of the location of the large source regions in the inventories, but have also revealed likely underestimates in the magnitude of their emissions, especially in the Northern Hemisphere. A regional study²⁰ has highlighted the advantage of averaging data to reveal smaller, localized NH₃ point sources, for which there is a single discernible source of pollution. Here we capitalize on nine years of IASI (Infrared Atmospheric Sounding Interferometer) measurements to produce a global distribution of NH₃ at hyperfine resolution to identify,

categorize and quantify the world's main NH₃ emission hotspots, down to the point source, and to benchmark the state-of-the-art emission inventory EDGAR (Emissions Database for Global Atmospheric Research) v4.3.1¹⁰.

Using an oversampling approach that exploits the variable spatial extent and coverage of the satellite pixels (Methods), a nine-year global average of daily cloud-free IASI observations was obtained. The global map is shown in Fig. 1a, along with three zooms over South and North America (Fig. 1b), Europe, northern Africa and Middle East (Fig. 1c), and Asia (Fig. 1d). We note that the global map shown in Fig. 1a differs from the first reported global distribution obtained from one year of IASI measurements¹³ in several aspects (in addition to the different periods considered for averaging) owing to major improvements in the retrieval algorithm, which takes into account the variable measurement sensitivity between regions^{15,21}. We analysed the map and isolated and identified 248 hotspot locations (Methods; Extended Data Figs. 1, 4; Supplementary Information). These consist of areas of limited geographical extent (<50 km) that exhibit a large local NH₃ enhancement and typically contain one or more closely located point sources. Hotspots were included in the list only if they could be identified unambiguously on the basis of satellite data alone (Methods). 178 regions with enhanced NH₃ columns, but with no clear, well-defined hotspots, were also inventoried (Extended Data Fig. 1; Supplementary Information). These source regions correspond to, for example, crop fields, biomass-burning areas, mixed sources, larger hotspots or several neighbouring ones. The largest regions are the Indo-Gangetic Plain, North China Plain and the biomass-burning-dominated West Africa and Amazonia. Regions dominated by biomass burning were mostly excluded from this study because identification of hotspots in such regions is very difficult and because we compare our results with a static-emission inventory that does not include emissions from fires (Methods). The hotspots were further analysed to determine the predominant emitter.

By combining information from visible imagery, publicly available inventories^{22,23} and online sources (Methods), the identified hotspots could be classified in three classes: agricultural, industrial and natural. Illustrative examples are shown in Fig. 2 and detailed figures are provided in the Supplementary Information.

The 83 hotspots in the agricultural class were consistently found to be associated with intensive animal farming, either in the form of open feedlots or within enclosed housings, as identified by aerial photographs. For instance, the localized NH₃ maximum found over Eckley-Yuma (Colorado, USA; Fig. 2a) can be seen to coincide spatially with two large cattle feedlots. These are situated in a large agricultural region dominated by centre-pivot-irrigated fields, but with much lower average NH₃ concentrations. Bakersfield and Tulare (California, USA) and Torreón (Mexico) are other examples of feedlot-dominated hotspots within a much larger intensive-agriculture region. Milford (Utah, USA), which is located in an otherwise remote mountainous area, is the home of massive pig farms, with associated open waste pits (Fig. 2b).

¹Université libre de Bruxelles (ULB), Service de Chimie Quantique et Photophysique, Atmospheric Spectroscopy, Brussels, Belgium. ²LATMOS/IPSL, UPMC Université Paris-06, Sorbonne Universités, UVSQ, CNRS, Paris, France. ³These authors contributed equally: Martin Van Damme, Lieven Clarisse. *e-mail: martin.van.damme@ulb.ac.be; lclariss@ulb.ac.be

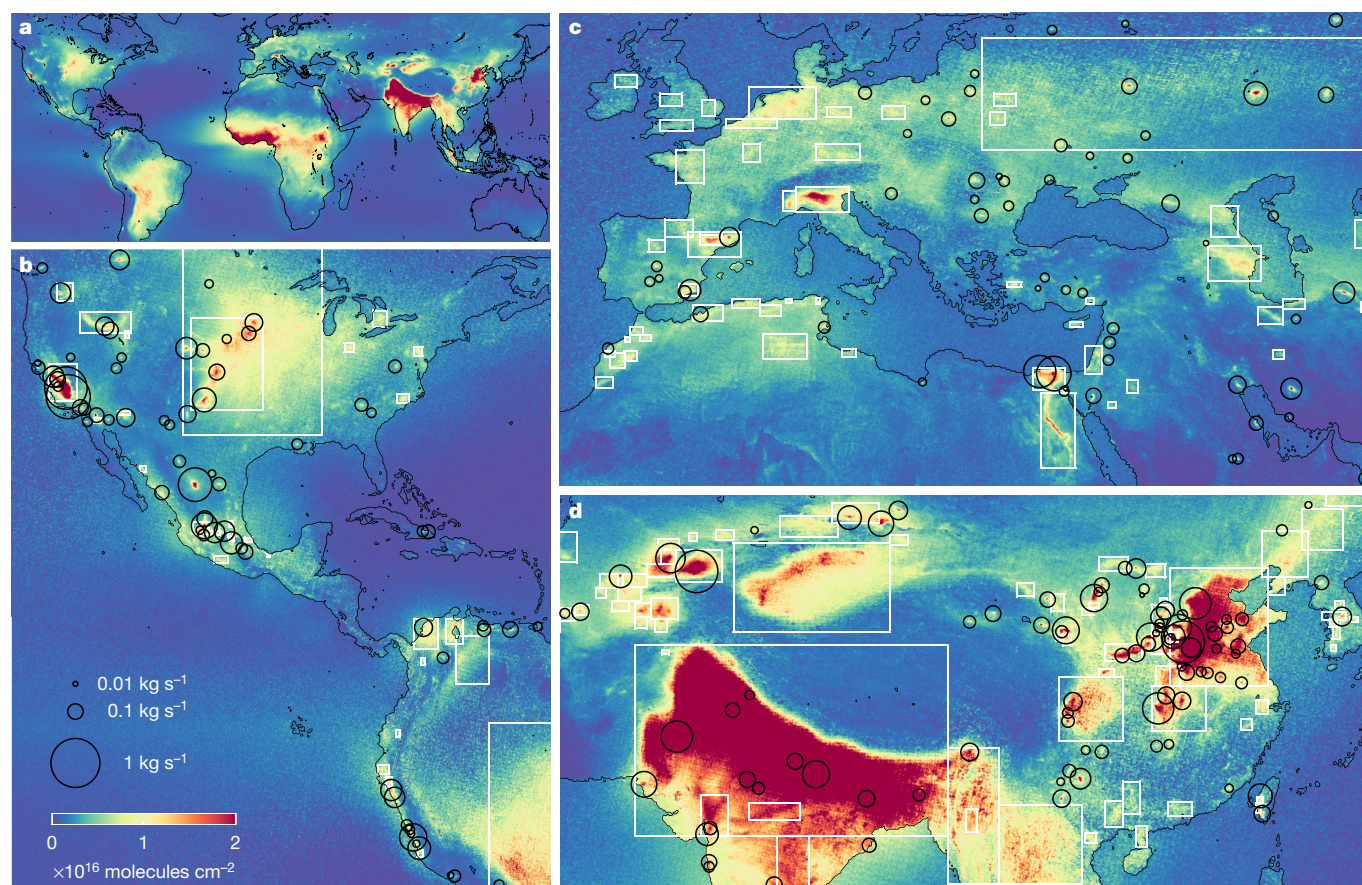


Fig. 1 | IASI nine-year oversampled average, hotspots and source regions. **a**, Nine-year global IASI average NH_3 distribution (in molecules per square centimetre). **b–d**, Zoom-ins over South and North America (**b**), Europe, northern Africa and the Middle East (**c**) and Asia (**d**). Hotspots

are indicated with black circles; their size quantifies the satellite-based emission fluxes (in kilograms per second). Source regions are delineated in white. The largest average NH_3 column is found over the Indus Valley (Pakistan) with a value of 1.1×10^{17} molecules cm^{-2} .

NH_3 emissions associated with poultry housings were also identified, for example, in the Alto Laran district (Peru; Fig. 2c) or in Basmakci (Turkey), a centre of egg production²⁴.

The second class (158 hotspots in total), that of industrial emitters, was in majority traced back to plants producing NH_3 -based fertilizer, for which over 130 sites were found. Well-isolated examples include the plants in Marvdasht (Iran; Fig. 2d), Pingsongxiang (Shanxi, China), Cherkasy (Ukraine), Sur Industrial Estate (Oman) and Beech Island (South Carolina, USA). Fertilizer plants are often found to be geographically close to their (agricultural) distribution market, such as the plants in the Ferghana Valley (Uzbekistan) and in the Nile Delta (Talkha and Abu Qir, Egypt). These industrial point sources clearly emerge in the nine-year IASI average, despite the already large background concentrations. Many fertilizer plants are also found near coal-related industries (coal mining, thermal power plants, coke production and other chemical coal industries). Secunda (South Africa) is an archetype of such a hotspot. In China, such examples are abundant; for instance, the large industry park in Shizuishan (Ningxia; Fig. 2e) or the intense hotspot over Zezhou-Gaoping (Shanxi). For these sites, fertilizer production is only part of the total industrial source. Other fainter industrial hotspots were found over nickel–cobalt mines (Moa and Nicaro, Cuba; Fig. 2f), soda-ash plants (Stuparei, Romania; Fig. 2g) and a complex of geothermal power plants (The Geysers, California, USA; Fig. 2h).

The third class includes natural emitters. Natural emissions from oceans, non-agricultural soils and plants represent a substantial part of the total atmospheric NH_3 budget⁷. However, these sources are generally too diffuse to appear as hotspots in the satellite data. Of all the hotspots that were identified, only the one near Lake Natron (Tanzania) is likely to have a natural origin. This hotspot occurs over Natron's main mudflat and may be related to the decay of algae. It is, however,

unclear why NH_3 emissions are larger at Lake Natron than at other soda lakes with similar regularly exposed mudflats. Other known natural point sources include seal and seabird colonies and volcanoes^{25–27}. Bird colonies are found in coastal areas and especially at high latitudes. In these areas, satellites are generally less effective as a detection and monitoring tool because of high turbulent mixing (which does not allow NH_3 to build up), high cloud cover and low thermal contrast (Methods). Enhanced NH_3 columns near some volcanoes were found to be associated with fires.

To assess the importance of the different point sources quantitatively, the nine-year-averaged emission fluxes were calculated for all hotspots and source regions using an inverse modelling approach (Methods). We compared these with the bottom-up emission inventory EDGAR (excluding biomass-burning regions, as detailed in Methods, Supplementary Information and Fig. 1). To test the validity of our approach, flux estimates from the identified source regions were first compared with EDGAR and are presented in Fig. 3 (diamonds). For 78% of the source regions, the fluxes agree within a factor of 3 (89% within a factor of 5) and, importantly, when all regions are considered, no major bias is observed.

For the flux calculations of the hotspots, background concentrations were removed to isolate the contribution of the point sources within (Methods). The flux comparisons are presented in Fig. 3, for agricultural (circles) and industrial (triangles) sources. It is immediately clear that emissions from almost all identified hotspots are underestimated in EDGAR, irrespective of their class. Of the 241 industrial and agricultural hotspots, only 7% agree within a factor of 2 and only one-third within one order of magnitude. 77 hotspots have a nearby local maximum in the inventory and can therefore be considered to be known, albeit with a flux rate that is too low. Representative examples

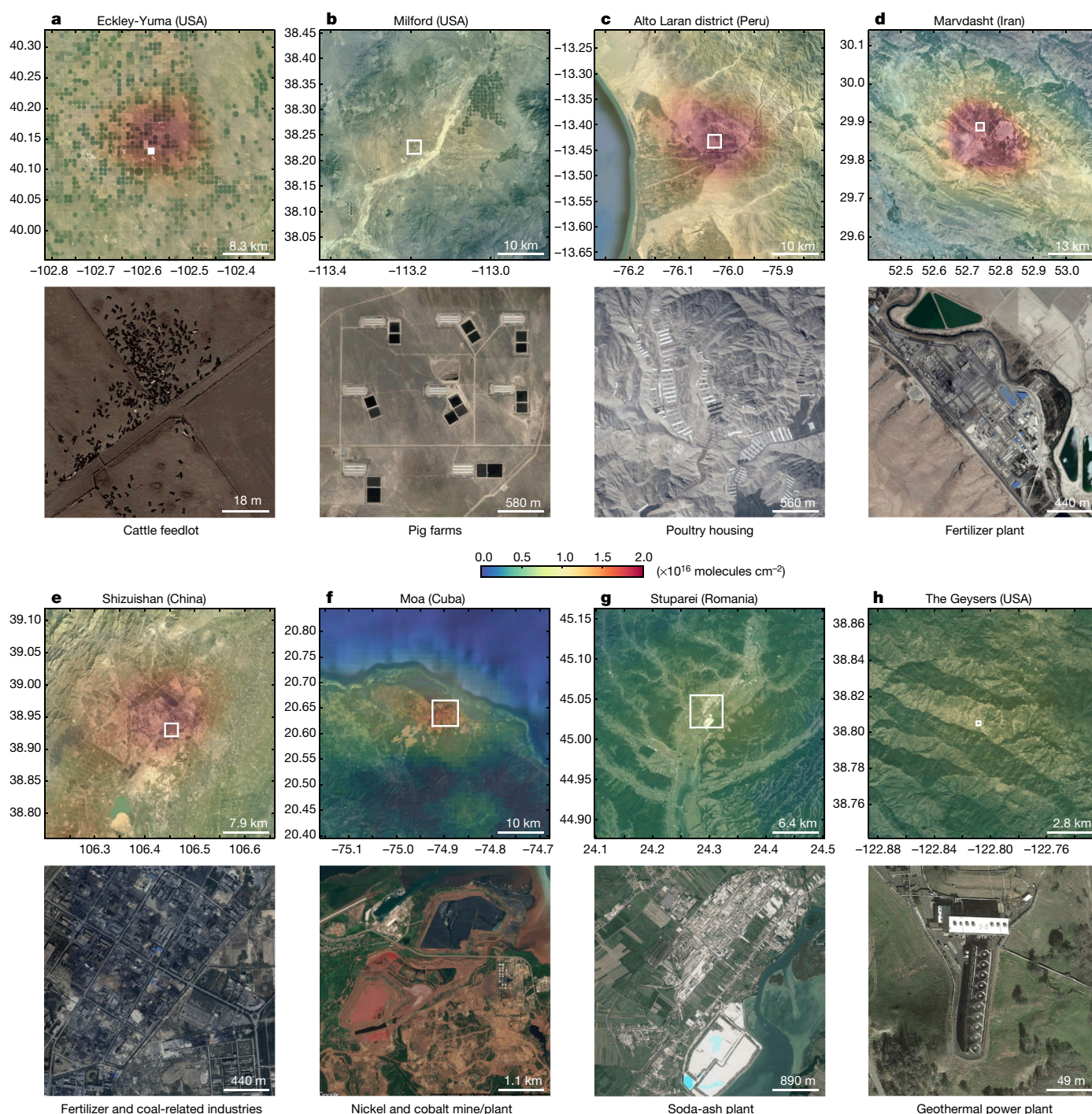


Fig. 2 | Examples of industrial and agricultural point sources.

a–h, For each site, the top panels overlay NH_3 total columns from the nine-year IASI average (in 10^{16} molecules cm^{-2}) on visible imagery (the vertical and horizontal axes correspond to latitude and longitude, respectively; the location of each site is also indicated in Extended Data Fig. 1 and Supplementary Fig. 1). The bottom panels offer a close-up

view of the areas delineated in white. The colour scale ranges from 0 to 2×10^{16} molecules cm^{-2} except for Shizuishan where it ranges from 0 to 3×10^{16} molecules cm^{-2} . The NH_3 source is indicated under each aerial photograph. Map data from Google Earth, CNES/Airbus, DigitalGlobe and Landsat/Copernicus.

shown in Fig. 3 include Sur Industrial Estate and Cherkasy for industry and Tulare and Torreón for agriculture. Many of the (industrial) point sources seem slightly displaced, by at least one EDGAR grid cell, from the identified hotspot centre, which corresponds to about 10 km (for example, Secunda and Beech Island). Some are so far away that they could not be included in the above count—notably the Marvdasht plant in Iran, which seems to be recorded in the inventory at a location over 20 km southwest from the actual place of emission. The other 164 hotspots do not represent a local maximum in EDGAR and are largely

underestimated compared to IASI data. There is one agricultural site (Chino, California, USA) and 69 industrial sites that can be considered completely absent from the EDGAR inventory because their fluxes are at least two orders of magnitude lower than the measurement.

For most of the hotspots, it was possible to calculate yearly emission fluxes and observe their time evolution over the nine years of IASI measurements. It is noteworthy that the onset or the discontinuation of anthropogenic activity could be detected unambiguously (Fig. 4). Numerous new industrial point sources that emerged within the IASI

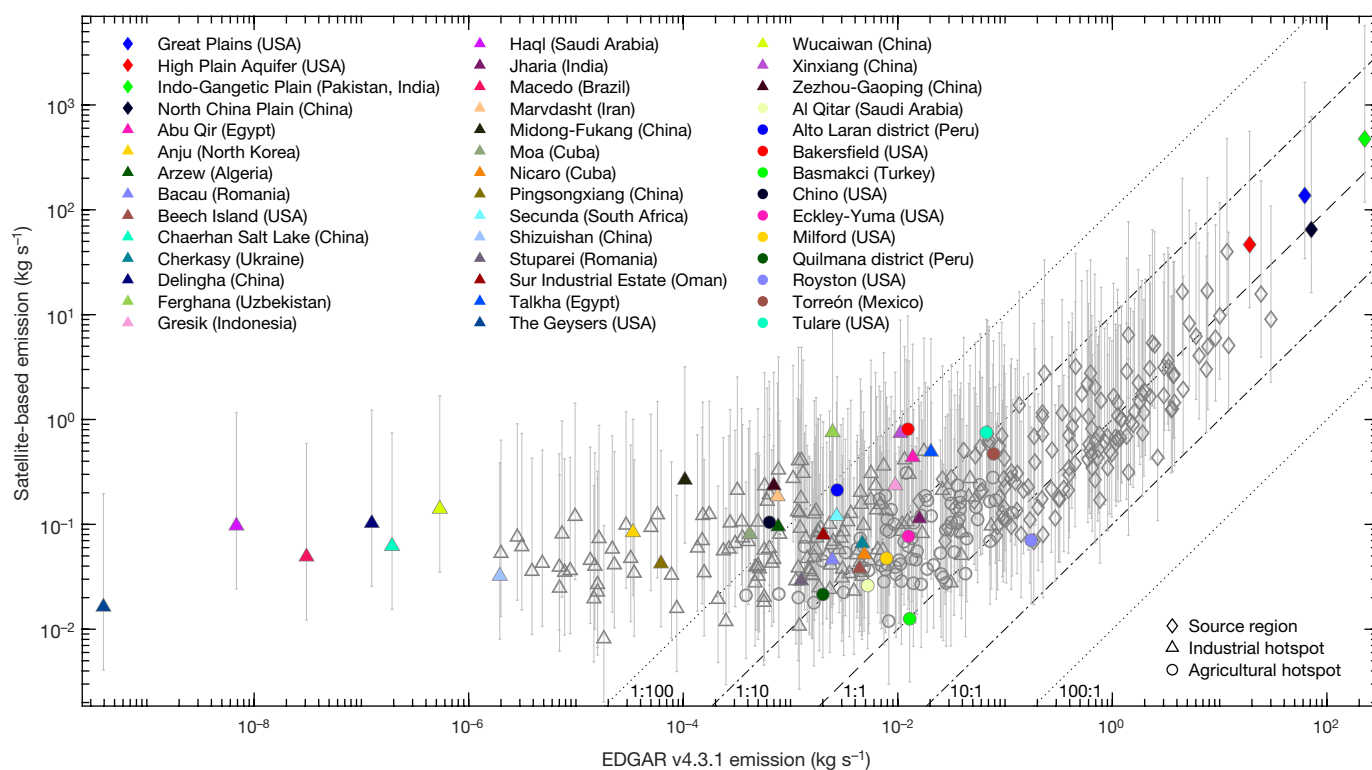


Fig. 3 | Satellite-derived emission fluxes versus a bottom-up emission inventory. Satellite-based emission estimates (in kilograms per second) for industrial (triangles) and agricultural (circles) hotspots and source regions (diamonds), compared with the EDGAR v4.3.1 emission inventory. The dashed, dash-dotted and dotted black lines represent ratios of EDGAR emission to satellite-based emission of 1:1, 1:10 or 10:1, and 1:100 or

100:1, respectively. The coloured symbols are for selected sites. Fluxes are calculated assuming a baseline NH_3 lifetime of 12 h (Methods); the error bars correspond to upper- and lower-bound flux estimates based on a lifetime of 1 h and 48 h, respectively. Biomass-burning regions are omitted from this comparison because EDGAR does not include emissions from fires.

measurement period were found in this way, especially in Asia. NH_3 is observed, for example, above Wucuiwan (Xinjiang, China) from 2012 onwards, which coincides with the establishment of a fertilizer factory. Industrial plant closures were also detected, for example, over Bacau (Romania) and Nicaro (Cuba) in 2013. Agriculture in transition was also identified, such as the rapid expansion of poultry farms in the central coastal regions of Peru (for instance, in the Alto Laran district).

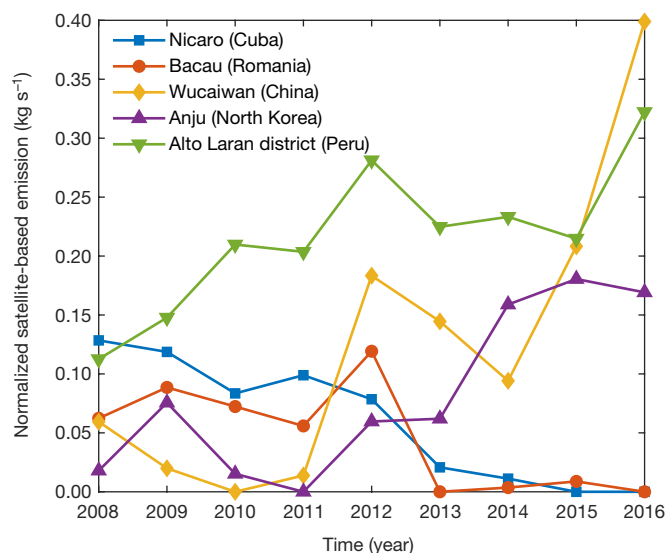


Fig. 4 | Examples of satellite-based ammonia emission trends. Time series of yearly averaged emission fluxes (in kilograms per second) over five industrial and agricultural point sources. The Nicaro, Bacau, Wucuiwan and Alto Laran district hotspots are discussed in the text; Anju illustrates the increase of fertilizer production in North Korea.

These examples show that IASI is capable of tracking the time evolution of relatively small emitters.

We have presented a detailed inventory of NH_3 hotspots using almost a decade of satellite measurements. Most hotspots are associated with either high-density animal farming or industrial fertilizer production. The prolonged measurement period made it possible to reveal all sources with a sustained emission history, including those that are either too weak or too small to be identified on shorter timescales. Satellite flux estimates for larger source regions agree with emission inventories within a factor of three, in contrast to strongly localized point sources, over which extremely large flux differences are found, irrespective of their origin. In addition, two-thirds of all the hotspots apparently do not have a corresponding local maximum in the EDGAR inventory. Our results suggest that it is necessary to revisit the input data used for traditional bottom-up NH_3 inventories and that space measurements can make an important contribution to the monitoring of ammonia emissions.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0747-1>.

Received: 6 February 2018; Accepted: 11 October 2018;
Published online 5 December 2018.

1. Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D. & Pozzer, A. The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature* **525**, 367–371 (2015).
2. Bauer, S. E., Tsigaridis, K. & Miller, R. Significant atmospheric aerosol pollution caused by world food cultivation. *Geophys. Res. Lett.* **43**, 5394–5400 (2016).
3. Galloway, J. et al. The nitrogen cascade. *Bioscience* **53**, 341–356 (2003).
4. Bobbink, R. et al. Global assessment of nitrogen deposition effects on terrestrial plant diversity: a synthesis. *Ecol. Appl.* **20**, 30–59 (2010).
5. Paerl, H. W., Gardner, W. S., McCarthy, M. J., Peierls, B. L. & Wilhelm, S. W. Algal blooms: noteworthy nitrogen. *Science* **346**, 175 (2014).

6. Shindell, D. T. et al. Improved attribution of climate forcing to emissions. *Science* **326**, 716–718 (2009).
7. Sutton, M. A. et al. Towards a climate-dependent paradigm of ammonia emission and deposition. *Phil. Trans. R. Soc. B* **368**, 20130166 (2013).
8. Reis, S., Pinder, R. W., Zhang, M., Lijie, G. & Sutton, M. A. Reactive nitrogen in atmospheric emission inventories. *Atmos. Chem. Phys.* **9**, 7657–7677 (2009).
9. Behera, S., Sharma, M., Aneja, V. & Balasubramanian, R. Ammonia in the atmosphere: a review on emission sources, atmospheric chemistry and deposition on terrestrial bodies. *Environ. Sci. Pollut. Res. Int.* **20**, 8092–8131 (2013).
10. *Emission Database for Global Atmospheric Research (EDGAR)*, release version 4.3.1 <http://edgar.jrc.ec.europa.eu/overview.php?v=431> (2016).
11. Huang, X. et al. A high-resolution ammonia emission inventory in China. *Glob. Biogeochem. Cycles* **26**, GB1030 (2012).
12. Meng, W. et al. Improvement of a global high-resolution ammonia emission inventory for combustion and industrial sources with new data from the residential and transportation sectors. *Environ. Sci. Technol.* **51**, 2821–2829 (2017).
13. Clarisse, L., Clerbaux, C., Dentener, F., Hurtmans, D. & Coheur, P.-F. Global ammonia distribution derived from infrared satellite observations. *Nat. Geosci.* **2**, 479–483 (2009).
14. Shephard, M. W. et al. TES ammonia retrieval strategy and global observations of the spatial and seasonal variability of ammonia. *Atmos. Chem. Phys.* **11**, 10743–10763 (2011).
15. Van Damme, M. et al. Global distributions, time series and error characterization of atmospheric ammonia (NH₃) from IASI satellite observations. *Atmos. Chem. Phys.* **14**, 2905–2922 (2014).
16. Shephard, M. W. & Cady-Pereira, K. E. Cross-track infrared sounder (CrIS) satellite observations of tropospheric ammonia. *Atmos. Meas. Tech.* **8**, 1323–1336 (2015).
17. Warner, J. X. et al. Increased atmospheric ammonia over the world's major agricultural areas detected from space. *Geophys. Res. Lett.* **44**, 2875–2884 (2017).
18. Streets, D. G. et al. Emissions estimation from satellite retrievals: a review of current capability. *Atmos. Environ.* **77**, 1011–1042 (2013).
19. Zhu, L. et al. Constraining U.S. ammonia emissions using TES remote sensing observations and the GEOS-Chem adjoint model. *J. Geophys. Res. Atmos.* **118**, 3355–3368 (2013).
20. Van Damme, M. et al. Evaluating 4 years of atmospheric ammonia (NH₃) over Europe using IASI satellite observations and LOTOS-EUROS model results. *J. Geophys. Res. Atmos.* **119**, 9549–9566 (2014).
21. Whitburn, S. et al. A flexible and robust neural network IASI-NH₃ retrieval algorithm. *J. Geophys. Res. Atmos.* **121**, 6581–6599 (2016).
22. IFDC. *Worldwide Ammonia Capacity Listing by Plant*. Report No. IFDC-FSR-10 (IFDC, 2016).
23. *Worldwide Syngas Database* www.globalsyngas.org (2018).
24. Yasar, S., Orhan, H. & Erensayin, C. Examining the nutritional and production characteristics of egg-farms in Basmakci County in Turkey. *Worlds Poult. Sci. J.* **59**, 249–259 (2003).
25. Theobald, M. R. et al. Ammonia emissions from a Cape fur seal colony, Cape Cross, Namibia. *Geophys. Res. Lett.* **33**, L03812 (2006).
26. Riddick, S. et al. High temporal resolution modelling of environmentally-dependent seabird ammonia emissions: description and testing of the guano model. *Atmos. Environ.* **161**, 48–60 (2017).
27. Uematsu, M. et al. Enhancement of primary productivity in the western North Pacific caused by the eruption of the Miyake-jima Volcano. *Geophys. Res. Lett.* **31**, L06106 (2004).

Acknowledgements IASI is a joint mission of EUMETSAT (European Organisation for the Exploitation of Meteorological Satellites) and the Centre National d'Etudes Spatiales (CNES, France). The research in Belgium was funded by F.R.S.-FNRS and the Belgian State Federal Office for Scientific, Technical and Cultural Affairs (Prodex arrangement IASI.FLOW). L.C. is a Research Associate (Chercheur Qualifié) with the Belgian F.R.S.-FNRS. C.C. is grateful to CNES for scientific collaboration and financial support. We thank M. Zondlo for discussions and R. Astoreca for assistance with the identification of certain hotspots. We gratefully acknowledge the Nederlandse Organisatie voor toegepast-natuurwetenschappelijk onderzoek (TNO) for the LOTOS-EUROS data used in the uncertainty analysis presented in the Supplementary Information.

Reviewer information Nature thanks F. Boersma and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.V.D. and L.C. obtained the first hyperresolved maps of NH₃ and performed the identification, classification and quantification of the sources, wrote the manuscript and prepared the figures. L.C., M.V.D., S.W. and J.H.-L. were responsible for the development of the retrieval algorithm and the processing of the IASI NH₃ dataset. D.H. was responsible for the development of the forward model. C.C. and P.-F.C. contributed to the text and interpretation of the results and supervised the research.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0747-1>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0747-1>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.V.D. or L.C. **Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

The IASI instrument and NH₃ dataset. IASI is a passive remote-sensing instrument that operates in downward viewing geometry to measure the infrared radiation emitted by Earth and its atmosphere in the 645–2,760 cm⁻¹ spectral range. The data used in this study are derived from the IASI instrument onboard the Metop-A platform, which was launched in 2006 in a Sun-synchronous orbit with a mean local solar overpass time of 9:30 a.m. and 9:30 p.m.²⁸ IASI has an elliptical footprint on the ground, ranging from 12 × 12 km² at nadir, up to 20 × 39 km² at its outermost viewing angle of 48°. Only measurements from the morning orbit have been used here, as IASI is generally more sensitive at this time to the atmospheric boundary layer, owing to more favourable thermal conditions²⁹. The impact of the IASI overpass time on the derived fluxes was evaluated using a regional model and found to be of the order of 4% ± 8% (provided that the diurnal variability is represented realistically in the model; see the uncertainty analyses in Supplementary Information, Supplementary Figs. 2, 4 and Supplementary Table 2).

This work relies on the ANNI-NH₃-v2.1R-I retrieval product and associated 2008–2016 dataset, which is described in full in ref.³⁰. The retrieval algorithm is based on the calculation of a hyperspectral range index and its conversion to a NH₃ column (in molecules per square centimetre) via an artificial neural network²¹. By exploiting a broad spectral range, full advantage is taken of the hyperspectral characteristics of IASI^{15,31}. This method provides a full uncertainty budget but, in contrast to constrained approaches, no information on the vertical sensitivity of each measurement (averaging kernels). To be able to properly assess inter-annual variability and trends, the retrieval relies on the ERA-Interim reanalysis³² for its meteorological input data (temperature, pressure and water vapour profiles). Information on cloud coverage is provided for each IASI observation within the IASI level 2 product that is disseminated by EUMETSAT in near-real time³³. Only measurements with a cloud fraction below 10% are retained. We note that until 3 March 2010, cloud fractions were not provided for all observations, resulting in reduced data availability and noisier distributions before this date (this is noticeable, for example, in Extended Data Fig. 4, especially for Torreón, where the distribution is more homogeneous after 2010). As no reliable variable global NH₃ vertical profiles can be obtained from measurements and as the choice was made to keep the analyses free from model input, a fixed vertical profile was used for the retrieval. Although validation^{34,35} and cross-comparison²⁹ studies have already excluded the existence of substantial biases in the IASI NH₃ products, in the Supplementary Information we explicitly show that the choice of vertical profile cannot realistically affect the conclusions of this study (see the uncertainty analyses and Supplementary Fig. 4). Differences between columns derived with a fixed vertical profile (baseline) and columns derived using variable modelled profiles are of the order of 2% ± 24% on a global scale, but may be substantially larger for individual locations linked to regional differences in meteorological mixing and recirculation.

Apart from the NH₃ measurements provided by IASI, three other NH₃ satellite products exist, from the instruments TES¹⁴, AIRS³⁶ and CrIS¹⁶. The overall methodology applied in this study could readily be applied to the last two datasets. The TES data however, would be less suitable because the spatial coverage of TES is too limited. It should also be noted that the NH₃ datasets from the other instruments are based on retrieval algorithms that use a priori information. To obtain unbiased flux estimates, more sophisticated flux calculations (for example, full formal inversion using atmospheric models) would be required. If these datasets would be taken at face value, we speculate that the inventory of hotspots and source regions would not be substantially different. However, the resulting fluxes would probably be even larger than those reported here, as preliminary intercomparisons have shown that retrieved columns are typically higher, partly because all of these instruments make early-afternoon measurements.

The global nine-year IASI average. The most widely used methods for averaging scattered satellite data rely on interpolation or binning on a rectangular latitude–longitude grid with a cell size comparable to the spatial resolution of the instrument. For IASI this is typically 0.125°, 0.25°, 0.5° or coarser. However, such approaches only take into account the location of the centre of the observation and not its spatial extent. Oversampling methods¹⁸ allow us to obtain averages at a much higher resolution and have already proven to be very successful in the study of NO₂, SO₂, HCHO and CO sources^{37–45}. These approaches exploit the fact that the location, shape and orientation of the satellite footprint on the ground varies from one orbit to another, so that by sacrificing temporal information, additional information can be obtained on the spatial distribution. This is akin to tomographic reconstruction in which multiple two-dimensional measurements from different angles are used to obtain a single three-dimensional measurement. Here we have adopted an oversampling technique very similar to the one applied for the averaging of NO₂ data from OMI³⁷. A grid size of 0.01° × 0.01° was chosen over the area ranging from 70° S to 70° N. For each observation, the footprint on the ground was calculated for IASI⁴⁶ and, for computational reasons, approximated

as an ellipse on a rectangular latitude–longitude grid. All measurements that cover a given cell were included in the calculation of the cell-averaged value. To account for differences in the footprint size, the averaging was performed with a weight inversely proportional to the area of each footprint. In previous work on NO₂³⁷, the weight of each measurement, in addition to the area, includes another term that is related to the measurement error. Here such a term was not added because the dynamic range of both the NH₃ column and its associated error is so large that weighing with either the relative or absolute error biases the average considerably^{15,21,30}. However, obviously erroneous measurements were excluded with an outlier test: for each measurement, the mean and standard deviation of all measurements within a 1° × 1° box were calculated, and the measurement was considered an outlier if it was more than 10 standard deviations from the average. In total, 0.014% of the measurements were excluded in this way. As an illustration of the method, an average of two days of IASI NH₃ measurements is shown in Extended Data Fig. 2. Because of cloud cover and the varying satellite sensitivity, not all periods of the year are sampled equally. However, analysis with an alternative approach, in which each month is assigned the same weight (see the uncertainty analyses in the Supplementary Information and Supplementary Fig. 4), reveals that the effect on the global distribution is modest, with emission estimates 4% ± 7% lower in the monthly based average.

The oversampling approach exploits all the spatial information contained within the measurement, without making any assumptions on, for example, the smoothness of the measured data. This is particularly useful for averaging measured data of a short-lived species such as NH₃, which exhibits sharp gradients close to its sources. Extended Data Fig. 3 compares an oversampled average (right; 0.01° × 0.01°) with a more traditional binned average (left; 0.25° × 0.25°) for the Nile Delta and Valley using the entire nine-year dataset. The figure demonstrates the considerable increase achieved in spatial resolution. Of particular relevance to this study are the two point sources in the north of the Gulf of Suez (Ain Sukhna and Al-Adabiya), which can only be distinguished in the oversampled average. We note that no artificial smoothing was performed in any of the maps that are shown here and in the Supplementary Information; their smooth appearance is solely due to the applied methodology, the sensitivity of the NH₃ algorithm and the availability of a large amount of data. For future work, one direction that could be envisaged is explicitly accounting for surface winds in the averaging. In certain cases, this could facilitate the isolation and identification of the hotspots in larger source regions⁴⁵.

Identification and attribution. The identification of hotspot locations and source regions was carried out manually. Although automated ways benefit from more consistency, no satisfactory set of criteria was found that could be applied globally or even per continent. This is particularly related to the fact that in certain areas, NH₃ concentrations are much more variable or NH₃ measurements are noisier. NH₃ emitted by biomass burning can especially hamper the identification of hotspots over a large area, even on a nine-year average (for example, most of South America and western Africa). As the static EDGAR emission inventory does not include biomass-burning emissions, the choice was made to completely exclude large regions with frequent biomass-burning episodes, both for the analysis of the hotspots and the analysis of source regions. Such regions were selected on the basis of the MODIS fire product⁴⁷; regions with more than 2 × 10⁻² fires per square kilometre for the entire period 2008–2016 were excluded. An assessment of specific fire emission inventories using IASI-derived NH₃ emissions can be found in earlier work^{48,49}. For the manual identification of hotspots, local maxima were identified that exhibited a characteristic smoothness compatible with a single source or a cluster of closely located sources and with a magnitude clearly above background values. This led to hotspot identification of areas typically smaller than 50 km. When the position of the maximum could not be clearly identified because it was spread over a large region, the area was classified as a source region instead of a hotspot. Because the above analysis was carried out on a best-effort basis, it necessarily involved some level of arbitrariness. In addition, numerous hotspots inevitably eluded identification, especially within larger source areas. The choice was also made not to use third-party sources (reported hotspots in the literature derived from in situ measurements or modelled data), which could bias the analysis. For instance, after it emerged that IASI is able to detect a complex of geothermal power plants in the United States, we could confirm the detection of other geothermal power plant locations, like the well-documented one in the Mount Amiata area in Italy⁵⁰. Another example is the fertilizer plant in Campana, close to Buenos Aires (Argentina), which appears on lists of NH₃ producers, and over which a small hotspot was indeed discerned a posteriori. However, its magnitude is not much larger than the hundreds of other local maxima in the region. Such hotspots that could not be identified unambiguously from the satellite data alone have therefore not been included.

After the list of hotspots and source regions was established, they were further analysed to ascertain the dominant origin of the observed NH₃. First, visible imagery was studied with Google Earth using both its image overlay and historical

imagery features. Agricultural activities (especially crop fields) exist in almost all populated places. Very frequently, where these activities looked no different in or outside the hotspot area (that is, showed no apparent signs of high-intensity animal farming), industrial sites could be spotted. So just by studying the imagery an educated guess could be made on whether industrial activity was involved or not. The high spatial resolution of the distribution proved to be invaluable here, as in many cases it allowed to pinpoint the location of the source within a few kilometres. This was less the case for sites on or near the coast, where owing to stronger winds the location of the maximum NH_3 column was sometimes observed up to 10 km away from its likely source (for example, the Nicaro mine in Cuba).

The presence of fertilizer production plants could be easily confirmed in most of the industrial cases by using a variety of publicly available inventories, in particular the IFDC Worldwide Ammonia Capacity Listing by Plant²², the Worldwide Syngas Database²³ and others (www.ammoniaindustry.com, www.fert.cn). Identification of the industry type was most difficult in China, where the industry is still rapidly developing and many fertilizer plants have yet to be included in these inventories. In addition, as already mentioned in the main text, fertilizer plants are often found close to other coal-related chemical industries that also emit ammonia. As soon as the presence of a fertilizer plant could be confirmed, the source was tagged as 'fertilizer industry', even when other industrial sources were clearly present. When this was not the case, it was tagged as 'other industry'; this category included mainly chemical or steel industries^{51,52}, but also the aforementioned geothermal power plant and two nickel–cobalt mines. Two hotspots were found to be associated with fires in coalmines (near Abakan, Russia and near Jharia, India) and these were classified under 'other industry' as well.

For a handful of hotspots (6)—in particular, those located near the large cities of Mexico City (Mexico), Bamako (Mali) and Niamey (Niger)—no obvious point sources could be identified. Most of the increases observed over other megacities were found to be too diluted or mixed with sources from outside to be classified as hotspots (for example, Sao Paulo in Brazil and Shanghai in China).

In addition to the identification and source attribution, a name was assigned to each hotspot and region. This was usually the name of the nearest medium-size city or for the larger areas the name of the region, province or geographical area—whichever was deemed most appropriate. For the hotspots, the approximate coordinates of their centres were established, as well as the size of an imaginary box around them, outside which the observed concentrations fall back to background values. These data were then used for the calculation of the flux data, as outlined below. The source regions were all approximated as rectangles on the latitude–longitude grid, the coordinates of which were also recorded.

Emission flux calculations. Following previous studies on NH_3 emissions from fires⁴⁹ and SO_2 sources⁵³, emission fluxes (E) were calculated from the IASI NH_3 distributions with a box model that assumes stationarity and constant first-order loss terms, that is, $E = M/\tau$. Here, M is the total mass contained within the assumed box, that is, the sum of the measured masses in each $0.01^\circ \times 0.01^\circ$ grid cell. These are obtained directly from their surface area and the measured satellite average NH_3 column (in molecules per square centimetre). The effective lifetime or residence time τ of NH_3 within a given box is defined for such a model as⁵⁴

$$\frac{1}{\tau} = \frac{1}{\tau_{\text{out}}} + \frac{1}{\tau_{\text{chem}}} + \frac{1}{\tau_{\text{dep}}}$$

with τ_{out} , τ_{chem} and τ_{dep} the lifetimes associated with export, chemical loss and deposition, respectively. Disregarding transport out of the box, the lifetime is believed to be of the order of a few hours to a few days (see Supplementary Table 1). In this study, the effective lifetime was set to 12 h on the basis of the limited data available in the literature. Using a set lifetime is a limitation of the method, as deposition rates and chemical losses are highly variable. The lifetime of NH_3 depends on the presence of other pollutants, meteorological variables (water vapour, clouds and temperature), winds (atmospheric mixing and recirculation) and the NH_3 column itself. In addition, the bidirectional nature of NH_3 surface–atmosphere fluxes can have a large and highly variable contribution⁵⁵. For many of the hotspots, there can be substantial transport out of the box, in which case the effective lifetime is smaller than 12 h. Given the average size of the boxes, the IASI flux estimates derived in this way are more likely to be underestimated than overestimated. For specific remote sites, however, 12 h could be too short. Therefore, to account for these uncertainties, fluxes were also calculated using lifetimes of 1 h and 48 h to provide conservative upper- and lower-bound flux estimates; the corresponding values are shown as error bars in Fig. 3. In addition, it is worth noting that analyses of temporal variability in emissions mostly do not depend on NH_3 retrieval or inverse model assumptions and hence are better constrained than the absolute fluxes. This is illustrated, for instance, in Fig. 4. Although other, more sophisticated methods are available⁵⁶, which can also estimate the effective lifetime from the data, an advantage of the box model is that it does not rely on external wind data.

More importantly, it is also applicable to both clusters of point sources and source areas, for which most of the other methods cannot be used.

Emissions for the larger source areas were calculated directly according to $E = M/\tau$. For the hotspots, a background column was first subtracted from the NH_3 columns, so as to include only the emission fluxes of the point sources that are responsible for the hotspots. The background column was estimated as the 10th percentile of all the $0.01^\circ \times 0.01^\circ$ average columns in the box around the hotspot. Other choices, such as the 5th or 15th percentile, are equally defensible. In the Supplementary Information (see the uncertainty analyses and Supplementary Figs. 3 and 4), we show that these alternative choices alter the fluxes by $-18\% \pm 7\%$ (5th percentile) and $14\% \pm 5\%$ (15th percentile). The background correction is illustrated in the Supplementary Information for all the hotspots mentioned in the main text.

Emission fluxes were also calculated from the EDGAR v4.3.1 inventory^{10,57} to compare with our satellite-based emission estimates. For 2010, the inventory is provided at a resolution of $0.1^\circ \times 0.1^\circ$ and separated into 17 sectors. Because the inventory provides emissions, only minor manipulation of the data was needed. The inventory was first regridded to the $0.01^\circ \times 0.01^\circ$ resolution of the IASI averages. For the source regions, EDGAR fluxes were then summed over the entire box and over all sectors. For the hotspots, only the fluxes of the cells that contain the assumed dominant point sources were considered because the satellite fluxes were calculated to represent only these. Those cells were determined as the largest 10th percentile cells of all the $0.01^\circ \times 0.01^\circ$ averaged observed columns in the box. As can be seen in the six-panel figures in the Supplementary Information (pages 2–29; the cells considered are located inside the black contour), this gives a reasonable approximation of what cells should be taken into account. Then, depending on the assigned source category, emissions found in the selected grid cells were added to obtain an estimated emission total of the dominant point sources in the box. We note that emissions from industry included the following sectors: power industry, oil refineries, transformation industry, combustion for manufacturing, process emissions during production and application, and solid waste and wastewater.

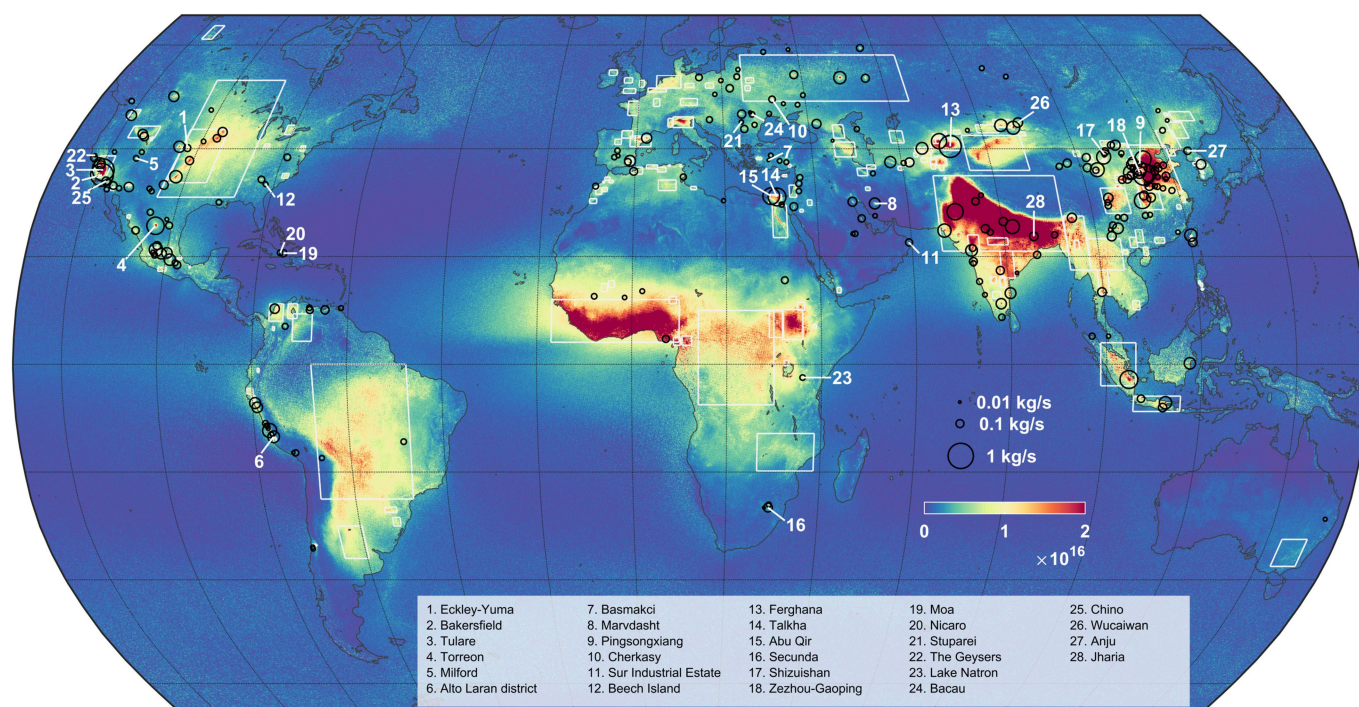
For several reasons, EDGAR may underestimate fluxes over hotspots even more than shown in Fig. 3. First, as explained above, the average effective lifetime is probably smaller than the 12 h used, leading to an underestimation of the IASI-derived flux (Supplementary Information). Second, IASI NH_3 measurements are more likely to underestimate NH_3 columns than to overestimate them because, depending on the thermal contrast, IASI (as any other infrared instrument) can be blind to the lower layer of the atmosphere, where NH_3 is emitted. Early validation efforts have also indicated that the IASI NH_3 measurements tend to underestimate ambient NH_3 concentrations, even when there is sufficient sensitivity^{34,35}. Third, in the methodology presented, background emissions are not removed for the flux calculations from EDGAR. For instance, for the flux estimate over a feedlot, the total flux still includes the contribution of other agricultural activities within the cells considered.

Data availability

The NH_3 map is available in NetCDF and KMZ formats. The latter file includes the identified hotspots and source regions and is provided in the Supplementary Information. The supplement also includes the 28 hotspot illustrations. The NetCDF file of the NH_3 map and the reanalysed IASI NH_3 dataset (ANNI- NH_3 -v2.1R-I) described in Methods are available from the PANGAEA repository (<https://doi.org/10.1594/PANGAEA.894736>); more recent versions of IASI NH_3 datasets are available from the AERIS data infrastructure (<http://iasi.aeris-data.fr>). The NH_3 product from IASI will also be operationally distributed by EUMETCast, under the auspices of the EUMETSAT Atmospheric Monitoring Satellite Application Facility (AC-SAF; <http://ac-saf.eumetsat.int>).

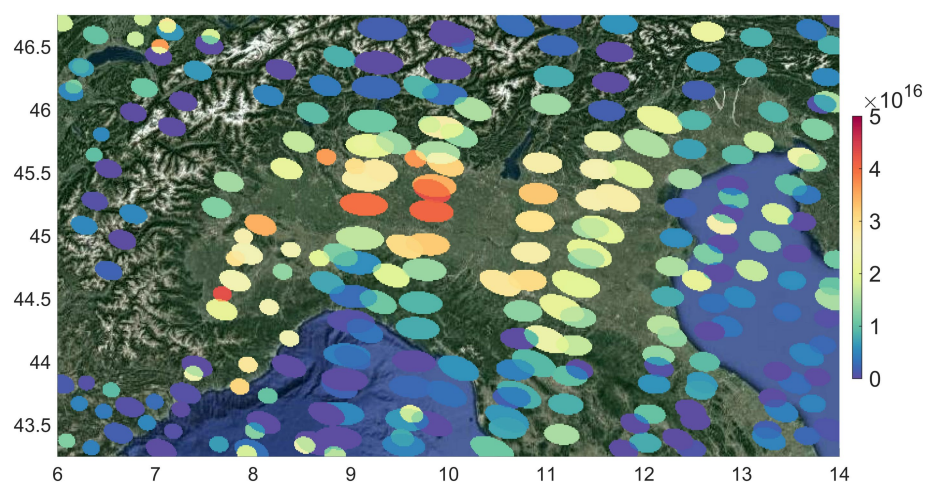
- Clerbaux, C. et al. Monitoring of atmospheric composition using the thermal infrared IASI/MetOp sounder. *Atmos. Chem. Phys.* **9**, 6041–6054 (2009).
- Clarisse, L. et al. Satellite monitoring of ammonia: a case study of the San Joaquin Valley. *J. Geophys. Res.* **115**, D13302 (2010).
- Van Damme, M. et al. Version 2 of the IASI NH_3 neural network retrieval algorithm: near-real-time and reanalysed datasets. *Atmos. Meas. Tech.* **10**, 4905–4914 (2017).
- Walker, J. C., Dudhia, A. & Carboni, E. An effective method for the detection of trace species demonstrated using the MetOp Infrared Atmospheric Sounding Interferometer. *Atmos. Meas. Tech.* **4**, 1567–1580 (2011).
- Dee, D. P. et al. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **137**, 553–597 (2011).
- August, T. et al. IASI on Metop-A: Operational Level 2 retrievals after five years in orbit. *J. Quant. Spectrosc. Radiat. Transf.* **113**, 1340–1371 (2012).
- Van Damme, M. et al. Towards validation of ammonia (NH_3) measurements from the IASI satellite. *Atmos. Meas. Tech.* **8**, 1575–1591 (2015).
- Dammers, E. et al. An evaluation of IASI- NH_3 with ground-based Fourier transform infrared spectroscopy measurements. *Atmos. Chem. Phys.* **16**, 10351–10368 (2016).

36. Warner, J. X., Wei, Z., Strow, L. L., Dickerson, R. R. & Nowak, J. B. The global tropospheric ammonia distribution as seen in the 13-year AIRS measurement record. *Atmos. Chem. Phys.* **16**, 5467–5479 (2016).
37. Wenig, M. O. et al. Validation of OMI tropospheric NO₂ column densities using direct-Sun mode Brewer measurements at NASA Goddard Space Flight Center. *J. Geophys. Res.* **113**, D16S45 (2008).
38. Fioletov, V. E., McLinden, C. A., Krotkov, N., Moran, M. D. & Yang, K. Estimation of SO₂ emissions using OMI retrievals. *Geophys. Res. Lett.* **38**, L21811 (2011).
39. Lu, Z., Streets, D. G., de Foy, B. & Krotkov, N. A. Ozone Monitoring Instrument observations of interannual increases in SO₂ emissions from Indian coal-fired power plants during 2005–2012. *Environ. Sci. Technol.* **47**, 13993–14000 (2013).
40. Pommier, M., McLinden, C. A. & Deeter, M. Relative changes in CO emissions over megacities based on observations from space. *Geophys. Res. Lett.* **40**, 3766–3771 (2013).
41. Zhu, L. et al. Anthropogenic emissions of highly reactive volatile organic compounds in eastern Texas inferred from oversampling of satellite (OMI) measurements of HCHO columns. *Environ. Res. Lett.* **9**, 114004 (2014).
42. de Foy, B., Lu, Z., Streets, D. G., Lamsal, L. N. & Duncan, B. N. Estimates of power plant NO_x emissions and lifetimes from OMI NO₂ satellite retrievals. *Atmos. Environ.* **116**, 1–11 (2015).
43. Fioletov, V. E., McLinden, C. A., Krotkov, N. & Li, C. Lifetimes and emissions of SO₂ from point sources estimated from OMI. *Geophys. Res. Lett.* **42**, 1969–1976 (2015).
44. Krotkov, N. A. et al. Aura OMI observations of regional SO₂ and NO₂ pollution changes from 2005 to 2015. *Atmos. Chem. Phys.* **16**, 4605–4629 (2016).
45. McLinden, C. A. et al. Space-based detection of missing sulfur dioxide sources of global air pollution. *Nat. Geosci.* **9**, 496–500 (2016).
46. Marsouin, A. & Brunel, P. *AAPP Documentation, Annex of Scientific Description, AAPP Navigation*. Report No. NWPSAF-MF-UD-005 (EUMETSAT, 2011).
47. Giglio, L., Schroeder, W. & Justice, C. O. The collection 6 MODIS active fire detection algorithm and fire products. *Remote Sens. Environ.* **178**, 31–41 (2016).
48. Whitburn, S. et al. Ammonia emissions in tropical biomass burning regions: Comparison between satellite-derived emissions and bottom-up fire inventories. *Atmos. Environ.* **121**, 42–54 (2015).
49. Whitburn, S. et al. Doubling of annual ammonia emissions from the peat fires in Indonesia during the 2015 El Niño. *Geophys. Res. Lett.* **43**, 11007–11014 (2016).
50. Bravi, M. & Basosi, R. Environmental impact of electricity from selected geothermal power plants in Italy. *J. Clean. Prod.* **66**, 301–308 (2014).
51. Wang, S. et al. Atmospheric ammonia and its impacts on regional air quality over the megacity of Shanghai, China. *Sci. Rep.* **5**, 15842 (2015).
52. Roe, S. M. et al. *Estimating Ammonia Emissions from Anthropogenic Nonagricultural Sources*. Draft Final Report (US Environmental Protection Agency, 2004).
53. Theys, N. et al. Volcanic SO₂ fluxes derived from satellite data: a survey using OMI, GOME-2, IASI and MODIS. *Atmos. Chem. Phys.* **13**, 5945–5968 (2013).
54. Jacob, D. J. *Introduction to Atmospheric Chemistry* (Princeton Univ. Press, Princeton, 1999).
55. Zhu, L. et al. Global evaluation of ammonia bidirectional exchange and livestock diurnal variation schemes. *Atmos. Chem. Phys.* **15**, 12823–12843 (2015).
56. de Foy, B., Wilkins, J. L., Lu, Z., Streets, D. G. & Duncan, B. N. Model evaluation of methods for estimating surface emissions and chemical lifetimes from satellite data. *Atmos. Environ.* **98**, 66–77 (2014).
57. Crippa, M. et al. Forty years of improvements in European air quality: regional policy-industry interactions with global impacts. *Atmos. Chem. Phys.* **16**, 3825–3841 (2016).



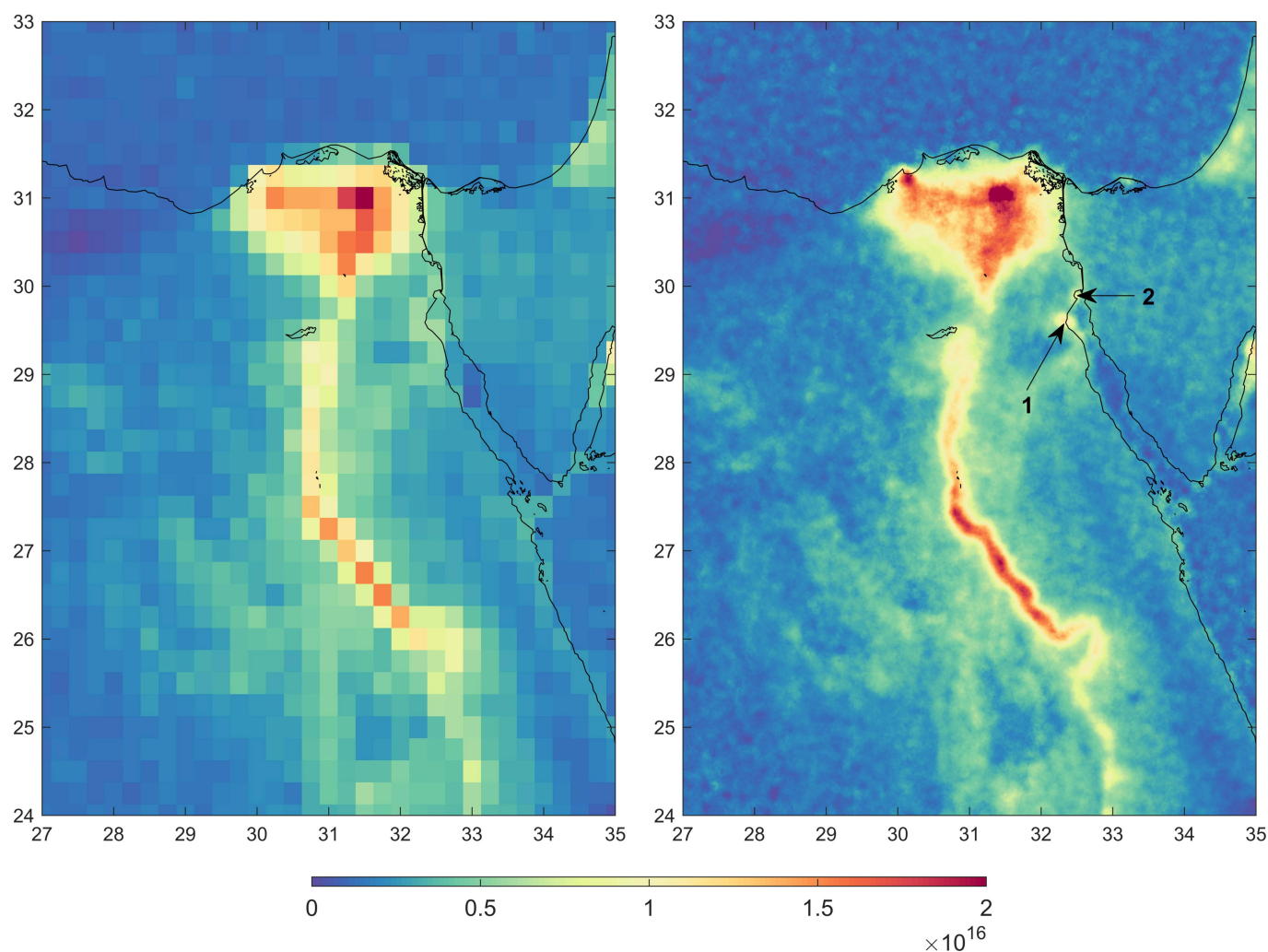
Extended Data Fig. 1 | Source areas and hotspot locations. Global nine-year NH_3 average (in molecules per square centimetre) with identified hotspots, their associated flux estimates (black circles), and source areas (white rectangles). In total 248 hotspots and 178 source areas

are indicated (see Supplementary Information for details). The locations and names of the hotspots discussed in the main text are also provided. The largest average NH_3 column is found over the Indus Valley (Pakistan) with a value of 1.1×10^{17} molecules cm^{-2} .



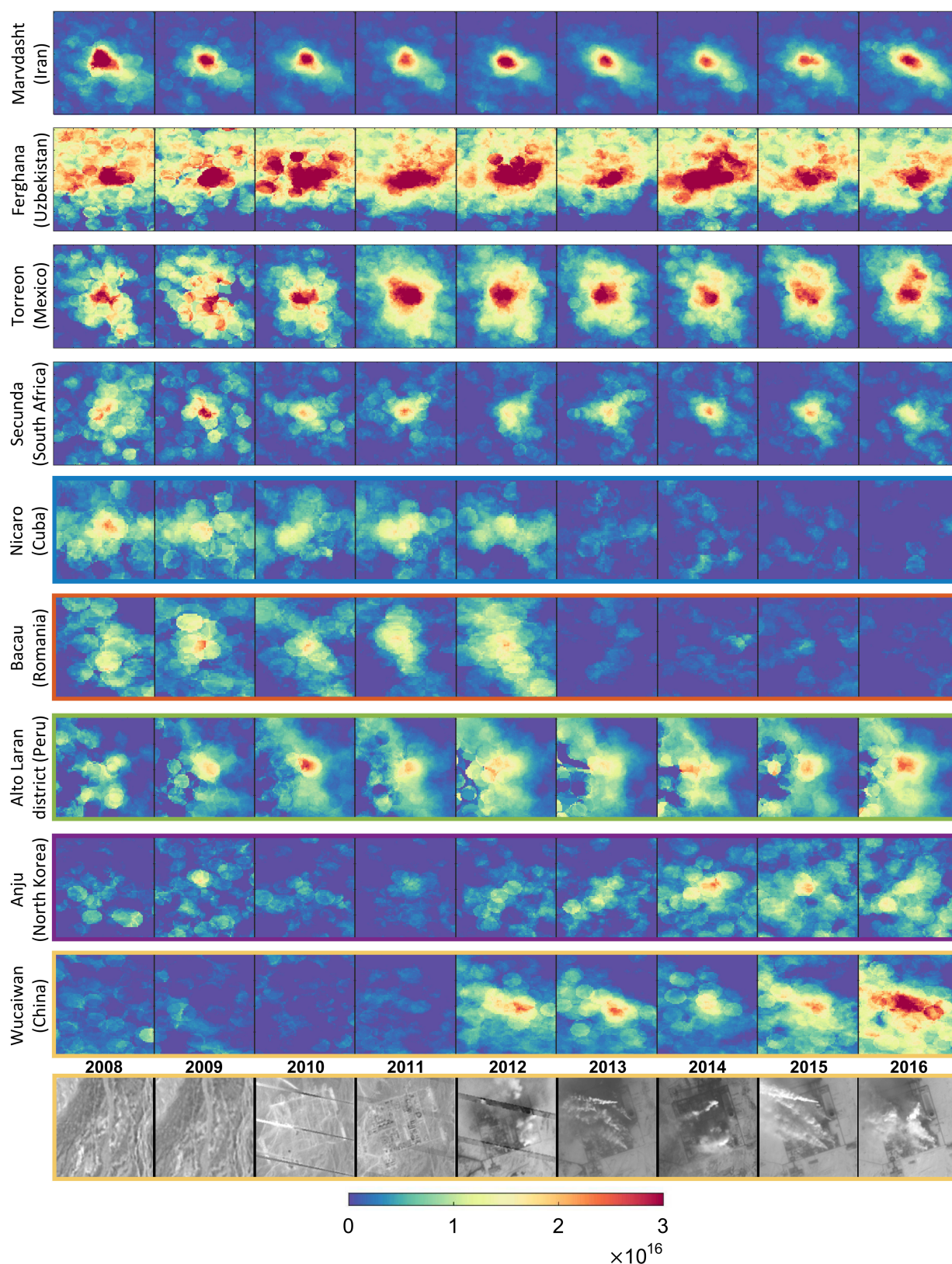
Extended Data Fig. 2 | Satellite footprint averaging. Example of two days (20 and 21 July 2016) of IASI/Metop-A morning NH₃ observations (in molecules per square centimetre) over the Po Valley. The elliptical

footprints of IASI are averaged on a $0.01^\circ \times 0.01^\circ$ high-resolution grid and weighted by the inverse of their footprint area. Map data from Google Earth and Landsat/Copernicus.



Extended Data Fig. 3 | Binned and oversampled averages over the Nile Delta and Valley. The right panel demonstrates the oversampled average (on a $0.01^\circ \times 0.01^\circ$ grid; maximum value of 3.1×10^{16} molecules cm^{-2}); the left panel shows the traditional binned average (on a $0.25^\circ \times 0.25^\circ$ grid; maximum value of 2.2×10^{16} molecules cm^{-2}) of IASI/Metop-A morning

NH₃ observations from 2008 to 2016 (in molecules per square centimetre; $12 \times 12 \text{ km}^2$ at nadir); see Methods for details. The increase in resolution allows the identification of two point sources in the north of the Gulf of Suez, Ain Sukhna (1) and Al-Adabiya (2), which cannot be singled out with a more traditional gridding approach.



Extended Data Fig. 4 | IASI NH_3 yearly distributions. Yearly NH_3 maps (in molecules per square centimetre) over several hotspots: (from top to bottom) Marvdasht (Iran), Ferghana (Uzbekistan), Torreón (Mexico), Secunda (South Africa), Nicaro (Cuba; blue in Fig. 4), Bacau (Romania; orange in Fig. 4), Alto Laran district (Peru; green in Fig. 4), Anju (North Korea; purple in Fig. 4) and Wucaiwan (China; yellow in Fig. 4).

The 2008 and 2009 distributions are noisier owing to reduced data availability (see Methods). The Wucaiwan industrial point source can be seen to appear in Landsat-5 (2008–2009), -7 (2010–2012) and -8 (2013–2016) images (bottom panels). Map data from Google Earth and Landsat/Copernicus.

Nonlinear rise in Greenland runoff in response to post-industrial Arctic warming

Luke D. Trusel^{1,2*}, Sarah B. Das², Matthew B. Osman³, Matthew J. Evans⁴, Ben E. Smith⁵, Xavier Fettweis⁶, Joseph R. McConnell⁷, Brice P. Y. Noël⁸ & Michiel R. van den Broeke⁸

The Greenland ice sheet (GrIS) is a growing contributor to global sea-level rise¹, with recent ice mass loss dominated by surface meltwater runoff^{2,3}. Satellite observations reveal positive trends in GrIS surface melt extent⁴, but melt variability, intensity and runoff remain uncertain before the satellite era. Here we present the first continuous, multi-century and observationally constrained record of GrIS surface melt intensity and runoff, revealing that the magnitude of recent GrIS melting is exceptional over at least the last 350 years. We develop this record through stratigraphic analysis of central west Greenland ice cores, and demonstrate that measurements of refrozen melt layers in percolation zone ice cores can be used to quantifiably, and reproducibly, reconstruct past melt rates. We show significant ($P < 0.01$) and spatially extensive correlations between these ice-core-derived melt records and modelled melt rates^{5,6} and satellite-derived melt duration⁴ across Greenland more broadly, enabling the reconstruction of past ice-sheet-scale surface melt intensity and runoff. We find that the initiation of increases in GrIS melting closely follow the onset of industrial-era Arctic warming in the mid-1800s, but that the magnitude of GrIS melting has only recently emerged beyond the range of natural variability. Owing to a nonlinear response of surface melting to increasing summer air temperatures, continued atmospheric warming will lead to rapid increases in GrIS runoff and sea-level contributions.

Melting across higher elevations of the GrIS results in liquid water infiltration, percolation, and either refreezing or storage within the porous firn layer. Such processes reduce ice-sheet surface albedo, increase firn temperatures⁷, and may generate impermeable ice layers that exacerbate ice-sheet runoff⁸, the proportion of surface melt leaving the ice sheet. Runoff across the margins of the GrIS is presently the leading source of mass loss from the ice sheet^{2,3,6}, and has been implicated in the centennial-scale slowdown of the overturning circulation in the North Atlantic Ocean⁹.

GrIS surface melting in 2012 was more expansive than at any time over the 40 years since we started measuring melt using satellites (in 1978)⁴. Proposed mechanisms driving melt in 2012 include contributions from anomalous radiative^{10–12} and non-radiative¹³ energy fluxes, and anticyclonic atmospheric circulation favouring advection of warm, dry air and clear sky conditions¹⁴. An ice core record from Summit Station (Fig. 1) demonstrated the exceptional nature of 2012 melt at high elevation (about 3,200 m), revealing¹⁵ that it last occurred at this site in 1889. At lower elevations, where melt occurs more frequently and at greater rates, there exist only limited ice-core-derived reconstructions of melt variability^{16–18} and no quantifiable ice-core-based reconstructions of melt intensity or runoff. Furthermore, large discrepancies among reanalysis products over Greenland before the mid-twentieth century^{6,19} limit their utility over longer timescales. Consequently, the true anomaly of recent melt intensity and runoff, and the longer-term evolution of these processes in response to climate forcing, remain poorly constrained.

Here we analyse refrozen melt layer stratigraphy in ice cores from central west Greenland (CWG) drilled between 2003 and 2015 to quantify annual melting over the last several centuries (Methods; Extended Data Fig. 1a). Significant co-variability between records from two cores separated by about 40 km in the western GrIS accumulation area (cores GC and D5; Fig. 1) indicates that regional melt, percolation and refreezing are well preserved and represented in these multi-century cores (Fig. 2 inset; Methods). We combine records from cores GC and D5 with a shallower core (core GW; Fig. 1, Extended Data Fig. 1b) to create a single 339-year stacked CWG record (Methods). CWG reveals frequent, but generally low-magnitude, melting during the last three centuries, with approximately 13-year periodicity and a marked departure from low-melt conditions over the last two decades (Fig. 2). We develop a complementary 364-year melt record using a fourth core (core NU) recovered from an ice cap on Nuussuaq Peninsula (Fig. 1) in 2015. The NU core record contains a greater overall magnitude of melt than the CWG combined core record owing to its lower elevation and coastal location (Extended Data Table 1), similar 13–16-year oscillations, and a distinct departure from baseline conditions in recent decades (Fig. 2). Common periodicities in the CWG and NU melt records suggest the influence of multiple known climate modes on melt variability (Extended Data Fig. 2; Methods). In all cores, there is also a clear pattern towards more frequent, thicker (and thus more intense) melt layers towards the present day (Extended Data Fig. 1c).

Our results show a pronounced 250% to 575% increase in melt intensity over the last 20 years, relative to a pre-industrial baseline period (eighteenth century) for cores NU and CWG, respectively (Fig. 2). Furthermore, the most recent decade contained in the cores (2004–2013) experienced a more sustained and greater magnitude of melt than any other 10-year period in the ice-core records. For GrIS cores, 2012 melt is unambiguously the strongest melt season on record. Both NU and CWG annual ice-core-derived melt records significantly ($P < 0.01$) correlate with one another over their 339 years of overlap, and both also with summer air temperatures from the Ilulissat region (Extended Data Table 2; Methods), relationships that improve after applying a 5-year moving average, probably reflecting the noise inherent to melt records owing to variability in meltwater percolation and refreezing^{8,16}. These empirically derived results revealing coherence between independent melt and temperature records emphasize broad-scale GrIS melt forcing, and suggest that summer warming (see Fig. 2) is an important component of the observed regional melt intensification.

GrIS ice cores provide a valuable point-scale view of melt history within the percolation zone, but records of past melt variability within the GrIS saturation and bare ice zones, where melt is more directly tied to mass loss and runoff, have remained elusive. To assess whether our ice-core-derived melt records also serve as quantitative archives of past regional-scale melt variability and intensity, we examine correlations between the CWG stack and recent modelled and observed melt parameters across the entire GrIS (Fig. 3; Methods). We find

¹Department of Geology, Rowan University, Glassboro, NJ, USA. ²Department of Geology and Geophysics, Woods Hole Oceanographic Institution, Woods Hole, MA, USA. ³Joint Program in Oceanography, Massachusetts Institute of Technology/Woods Hole Oceanographic Institution, Woods Hole, MA, USA. ⁴Department of Chemistry, Wheaton College, Norton, MA, USA. ⁵Polar Science Center, Applied Physics Laboratory, University of Washington, Seattle, WA, USA. ⁶Department of Geography, University of Liège, Liège, Belgium. ⁷Division of Hydrologic Sciences, Desert Research Institute, Reno, NV, USA. ⁸Institute for Marine and Atmospheric Research, Utrecht University, Utrecht, Netherlands. *e-mail: trusel@rowan.edu

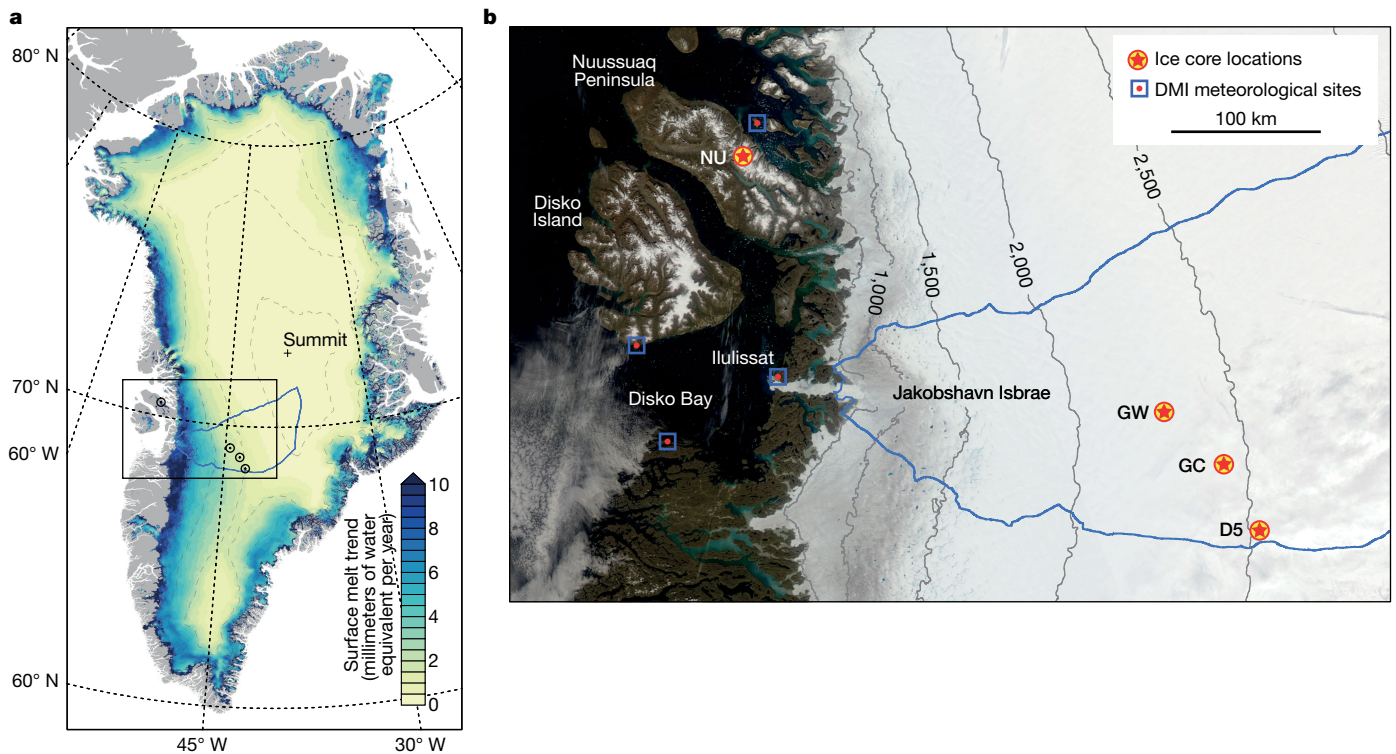


Fig. 1 | Positive trends in Greenland surface melt and locations of ice cores used to reconstruct past surface melt variability. **a**, Positive trends in annual surface meltwater production over 1958–2016, as simulated by RACMO2.3p2⁵. **b**, Locations of our ice cores situated within the Jakobshavn drainage basin (blue outline; basin 7.1) of CWG and on the

Nuussuaq Peninsula, as well as air-temperature observations integrated into a composite Ilulissat air-temperature record (Methods; Extended Data Table 2). The satellite image in **b** was obtained with the NASA MODIS instrument (from the NASA Rapid Response data archive).

wide-ranging significant ($P < 0.01$) positive correlations between the CWG melt record and RACMO2-modelled⁵ melt (Fig. 3a) and refreezing (Fig. 3b), as well as satellite-observed melt duration⁴ (Fig. 3d). Further supporting these results are significant ($P < 0.01$) relationships between our NU and CWG melt records and long-term pan-Greenland air temperature observations (Extended Data Table 3; Methods). Moreover, we find significant ($P < 0.01$) and wide-ranging correlations

between the ice-core-derived melt record and RACMO2-simulated meltwater runoff across the ice-sheet margins (Fig. 3c). These ice-core model relationships indicate that rates of meltwater production, refreezing and runoff across much of Greenland are all at multi-century highs, assuming that the spatial character of melt and high correlations between modelled GrIS-integrated melt, refreezing, and runoff are robust and have remained stationary through time (Extended Data Fig. 4).

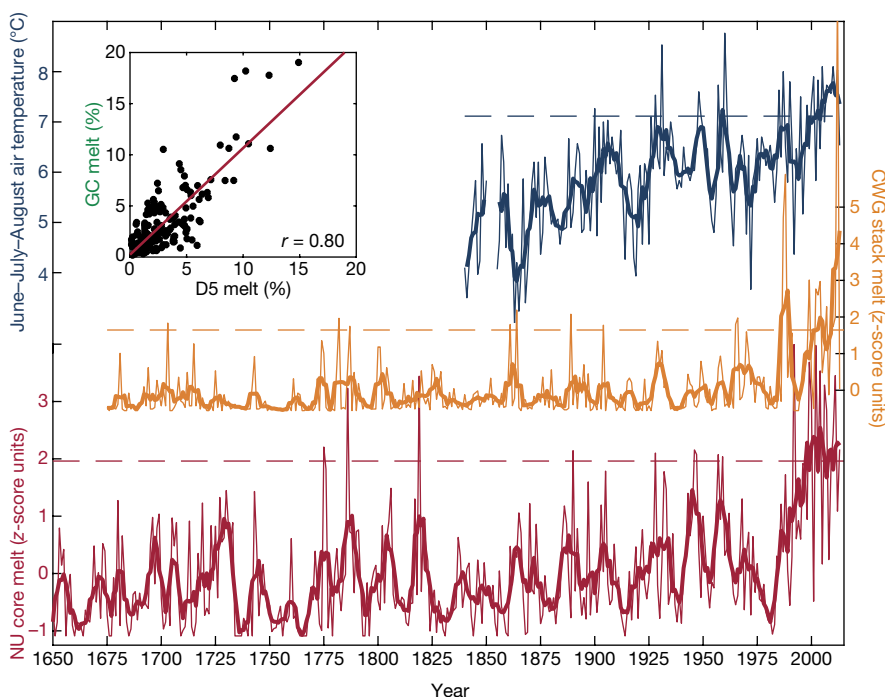


Fig. 2 | Multi-century evolution of CWG surface melt from ice cores. Observed summer (June–August) air temperatures from the Ilulissat area and melt intensity (in standardized z-score units) from the CWG ice-core-derived melt stack and from the NU core on Nuussuaq Peninsula (see Methods). Horizontal dashed lines show means over the last twenty years of the ice-core records (1994–2013). Bold lines show 5-year smoothed time series. Significant correlation ($P < 0.01$; $n^* = 106$) between 5-year running means of GC and D5 melt per cent (inset) reveals that variability in total annual surface melt, as well as broader regional melt variability, is well captured by our ice cores.

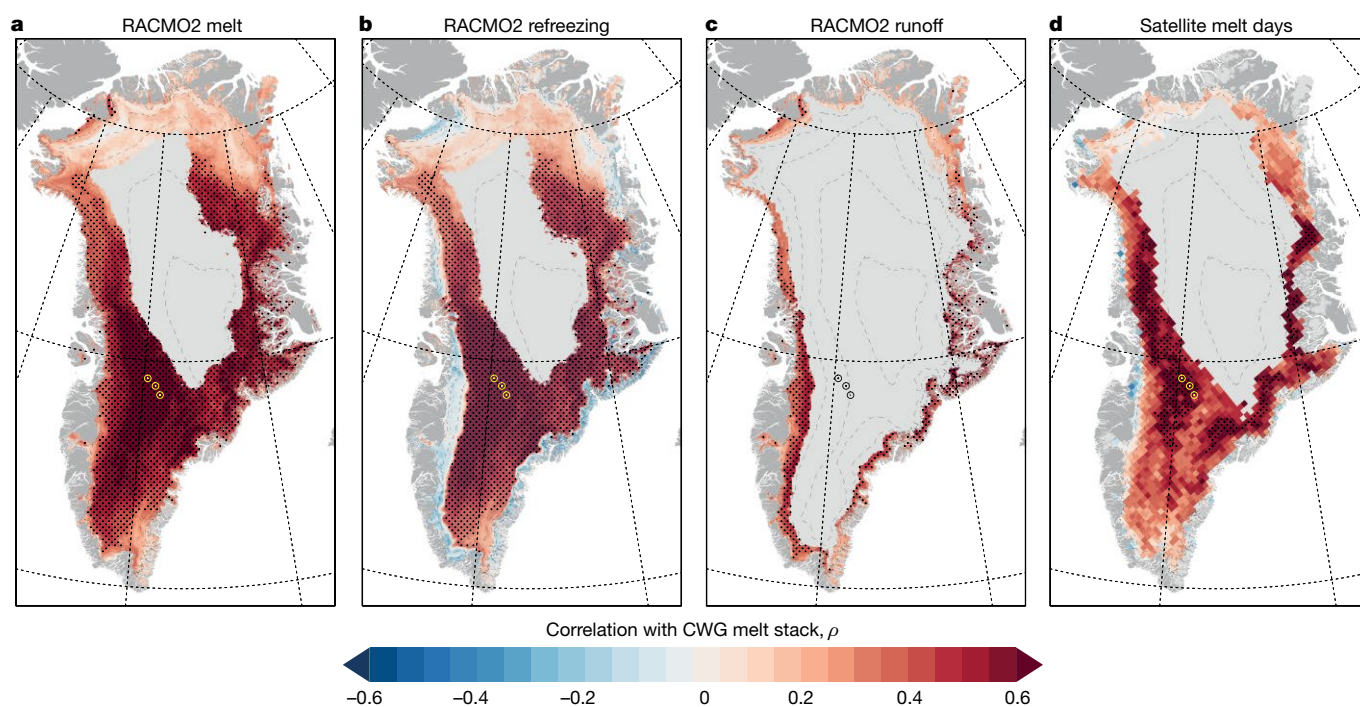


Fig. 3 | Spatially broad representation of melt processes captured by CWG ice cores. **a–d**, Spearman rank order correlations between the CWG melt stack and RACMO2-modelled annual melt (**a**), refreezing (**b**) and runoff (**c**) over 1958–2013⁵, and with satellite-observed melt over 1988–2013⁴ (**d**; Methods). Areas of significant correlation ($P < 0.01$) are

denoted by the black stipple pattern. Correlations are calculated only for years where the gridded data indicated melt for $>50\%$ of the years (28 years in **a–c**; 13 years in **d**). The location of cores used in the CWG stack is denoted by yellow (**a**, **b**, **d**) or black (**c**) points. Contour interval, 500 m.

Given the strong correlations between modern percolation zone melt and ice-sheet meltwater runoff, we produce the first observationally constrained record of ice-sheet runoff variability using a stepwise calibration/validation procedure between our ice-core melt records and RACMO2 runoff (Methods). Performing this analysis against integrated runoff from each GrIS surface drainage basin (Extended Data Fig. 5), as well as total GrIS runoff (Fig. 4a), we find that the combined CWG and NU melt records possess high predictive skill in reconstructing runoff over 1958–2013 across much of the GrIS (Extended Data Figs. 5, 6). Our results show that relatively stable and low runoff occurred before the 1990s in all individual basins (Extended Data Fig. 5) as well as across the broader ice sheet. Our reconstructions are consistent with modern model-based runoff estimates (Fig. 4a) and limited direct runoff measurements²⁰. Divergence in past runoff estimates derived from longer reanalyses^{6,19} hinders centennial-scale comparisons, but also highlights in particular the value of our new observationally based reconstruction farther back in time. We show that an exceptional rise in runoff has occurred over the last two decades, equating to an approximately 50% increase in GrIS-integrated runoff compared to pre-industrial runoff, and a 33% increase over the twentieth century alone (Fig. 4a).

The onset of industrial-era Arctic warming occurred in the mid-nineteenth century²¹ and differential smoothing analysis likewise indicates increases in GrIS runoff initiated shortly thereafter (Fig. 4b; Methods). The median onset of positive trends in GrIS runoff are also coincident with the median onset of weakening Atlantic meridional overturning circulation⁹. Emergence of runoff beyond the natural range of variability, however, has only very recently occurred (Fig. 4b). Although the precise mechanisms behind these rapid recent changes remain unconstrained, the high temporal resolution of our reconstruction allows us to explore a range of possibilities. Runoff emergence timing follows the emergence (that is, steep decline) of pan-Arctic summer sea ice (Methods; Fig. 4b, Extended Data Table 4), supporting prior hypotheses that Arctic sea ice loss may amplify GrIS melt and runoff²². The pronounced sea ice and runoff changes are also roughly coincident with a negative shift in the summer North Atlantic Oscillation, marking

enhanced summertime anticyclonic conditions over Greenland known to contribute to warming²³ and a positive melt–albedo feedback¹². Sea-ice loss is also suggested to be more directly responsible for atmospheric blocking over Greenland, thus contributing to recent North Atlantic Oscillation trends²². Regional atmospheric²⁴ and sea-ice²⁵ trends may also represent a teleconnected pan-Arctic response to tropical Pacific sea surface temperature forcing.

Our reconstruction quantifies the exceptional magnitude of present-day melt and runoff relative to the last several centuries. Their rapid intensification over the last two decades also illustrates a clear non-linear melt–temperature relationship (Fig. 4c; Methods). Similar late-twentieth-century melt acceleration was found using records from an Antarctic Peninsula ice core²⁶, and attributed to a nonlinear response of melting to climate warming more broadly across Antarctica, owing largely to the melt–albedo positive feedback²⁷. At all of our core sites, 2012 melt was more intense than any other year according to two distinct reanalysis-forced regional climate models that extend back to 1958⁵ and 1979⁶ (Fig. 4c). Our ice-core results provide further context and reveal that these 2012 melt rates are exceptional highs for at least the past 350 years. If an air-temperature reconstruction from the nearby Canadian Arctic is regionally representative, GrIS melt and runoff experienced in the last decade is likely also to be unprecedented over the last 6,800–7,800 years²⁸. The nonlinear melt–temperature sensitivity also helps explain why episodes of mid-twentieth-century warmth resulted in less intense and less sustained melting compared to the last two decades, despite being only marginally cooler (Fig. 2). Additional factors, such as recent sea-ice losses, as well as regional and teleconnected general circulation changes, may also play a part in amplifying the melt response. Moreover, this melt–temperature nonlinearity indicates that only limited additional warming will greatly enhance the area of the ice sheet subject to meltwater runoff. Indeed, even high-elevation regions of the percolation zone, such as at our CWG core sites, have already warmed enough to induce the melt intensification and firn densification necessary to generate meltwater runoff in some years, as opposed to full melt refreezing and retention (Fig. 4c).

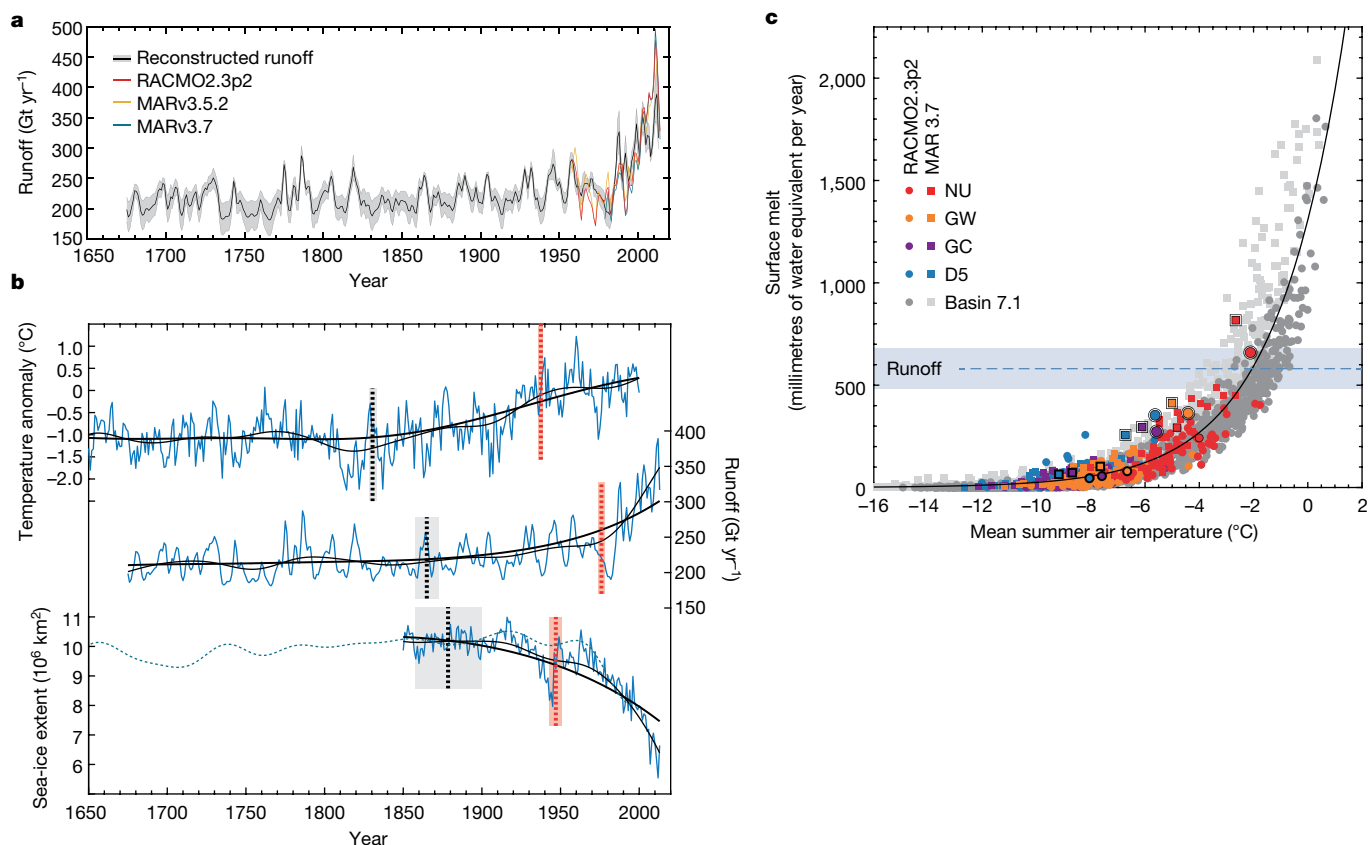


Fig. 4 | Exceptional rise in Greenland ice-sheet runoff and climate warming context. **a**, GrIS-integrated meltwater runoff, as simulated by regional climate models^{5,6} (coloured lines; 5-year smoothed) and reconstructed using the NU and CWG ice-core-derived melt records (black line; 95% confidence interval shaded; see Methods). **b**, Median onset of significant trends (vertical black dotted lines) and climate emergence above pre-industrial (vertical red dotted lines) for mean Arctic temperatures²¹ (top), our ice-core-derived runoff reconstruction (middle) and two summer Arctic sea-ice extent datasets^{29,30} (bottom; Methods). Median absolute deviations of trend onsets and climate emergence shown as shaded boxes. Thin and bold black lines denote 15-year and 50-year Gaussian smoothed series. **c**, Recent modelled evolution of mean summer (JJA) near-surface air temperature and surface

melt (in millimetres of water equivalent per year) across CWG. Ice core sites are shown as coloured points, and a Jakobshavn basin (basin 7.1; Fig. 1) elevational transect as grey points from RACMO2.3p2 (circles) and MARv3.7 (squares). Means over the past 20 years of the ice-core records (1994–2013) at core sites are denoted by points with single black border, and peak melting in 2012 by double black borders. The evolution of CWG ice-sheet melt in response to a warming climate is well represented by an exponential function (black curve). Recent melt rates at our percolation zone core sites approach conditions where the models have recently begun to simulate meltwater runoff (blue dashed line indicates mean runoff-linked melt rate and the shaded region corresponds to ± 1 s.d.; see Methods for details).

Our GrIS melt and runoff reconstructions highlight how ice-core-based surface mass balance reconstructions add important additional in situ context to the relatively brief instrumental, satellite and climate reanalysis eras. Today, surface melting and melt-induced runoff in Greenland occur at magnitudes not previously experienced over at least the last several centuries, if not millennia. Melt–temperature nonlinearity and general circulation changes mean that further twenty-first-century warming has important implications for the ice-sheet mass balance, by accelerating the intensity of surface melting and amplifying GrIS contributions to global sea-level rise.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0752-4>.

Received: 4 May; Accepted: 18 September 2018;
Published online 5 December 2018.

- Hanna, E. et al. Ice-sheet mass balance and climate change. *Nature* **498**, 51–59 (2013).
- Enderlin, E. M. et al. An improved mass budget for the Greenland ice sheet. *Geophys. Res. Lett.* **41**, 2013GL059010 (2014).
- van den Broeke, M. R. et al. On the recent contribution of the Greenland ice sheet to sea level change. *Cryosphere* **10**, 1933–1946 (2016).

- Tedesco, M. et al. Evidence and analysis of 2012 Greenland records from spaceborne observations, a regional climate model and reanalysis data. *Cryosphere* **7**, 615–630 (2013).
- Noël, B. et al. Modelling the climate and surface mass balance of polar ice sheets using RACMO2—Part 1: Greenland (1958–2016). *Cryosphere* **12**, 811–831 (2018).
- Fettweis, X. et al. Reconstructions of the 1900–2015 Greenland ice sheet surface mass balance using the regional climate MAR model. *Cryosphere* **11**, 1015–1033 (2017).
- Humphrey, N. F., Harper, J. T. & Pfeffer, W. T. Thermal tracking of meltwater retention in Greenland's accumulation area. *J. Geophys. Res.* **117**, F01010 (2012).
- Machguth, H. et al. Greenland meltwater storage in firn limited by near-surface ice formation. *Nat. Clim. Chang.* **6**, 390–393 (2016).
- Noël, B. et al. Anomalous weak Labrador Sea convection and Atlantic overturning during the past 150 years. *Nature* **556**, 227–230 (2018).
- Bennartz, R. et al. July 2012 Greenland melt extent enhanced by low-level liquid clouds. *Nature* **496**, 83–86 (2013).
- Van Tricht, K. et al. Clouds enhance Greenland ice sheet meltwater runoff. *Nat. Commun.* **7**, 10266 (2016).
- Hofer, S., Tedstone, A. J., Fettweis, X. & Bamber, J. L. Decreasing cloud cover drives the recent mass loss on the Greenland Ice Sheet. *Sci. Adv.* **3**, e1700584 (2017).
- Fausto, R. S. et al. The implication of nonradiative energy fluxes dominating Greenland ice sheet exceptional ablation area surface melt in 2012. *Geophys. Res. Lett.* **43**, 2649–2658 (2016).
- Hanna, E. et al. Atmospheric and oceanic climate forcing of the exceptional Greenland ice sheet surface melt in summer 2012. *Int. J. Climatol.* **34**, 1022–1037 (2014).

15. Keegan, K. M., Albert, M. R., McConnell, J. R. & Baker, I. Climate change and forest fires synergistically drive widespread melt events of the Greenland Ice Sheet. *Proc. Natl Acad. Sci. USA* **111**, 7964–7967 (2014).
16. Graeter, K. A. et al. Ice core records of West Greenland melt and climate forcing. *Geophys. Res. Lett.* **45**, 3164–3172 (2018).
17. Herron, M. M., Herron, S. L. & Langway, C. C. Climatic signal of ice melt features in southern Greenland. *Nature* **293**, 389–391 (1981).
18. Kameda, T. et al. Melt features in ice cores from Site J, southern Greenland: some implications for summer climate since AD 1550. *Ann. Glaciol.* **21**, 51–58 (1995).
19. van den Broeke, M. et al. Greenland ice sheet surface mass loss: recent developments in observation and modeling. *Curr. Clim. Change Rep.* **3**, 345–356 (2017).
20. Ahlström, A. P., Petersen, D., Langen, P. L., Citterio, M. & Box, J. E. Abrupt shift in the observed runoff from the southwestern Greenland ice sheet. *Sci. Adv.* **3**, e1701169 (2017).
21. Abram, N. J. et al. Early onset of industrial-era warming across the oceans and continents. *Nature* **536**, 411–418 (2016).
22. Liu, J. et al. Has Arctic sea-ice loss contributed to increased surface melting of the Greenland ice sheet? *J. Clim.* **29**, 3373–3386 (2016).
23. Fettweis, X. et al. Brief communication 'Important role of the mid-tropospheric atmospheric circulation in the recent surface melt increase over the Greenland ice sheet'. *Cryosphere* **7**, 241–248 (2013).
24. Ding, Q. et al. Tropical forcing of the recent rapid Arctic warming in northeastern Canada and Greenland. *Nature* **509**, 209–212 (2014).
25. Ding, Q. et al. Influence of high-latitude atmospheric circulation changes on summertime Arctic sea ice. *Nat. Clim. Chang.* **7**, 289–295 (2017).
26. Abram, N. J. et al. Acceleration of snow melt in an Antarctic Peninsula ice core during the twentieth century. *Nat. Geosci.* **6**, 404–411 (2013).
27. Trusel, L. D. et al. Divergent trajectories of Antarctic surface melt under two twenty-first-century climate scenarios. *Nat. Geosci.* **8**, 927–932 (2015).
28. Lecavalier, B. S. et al. High Arctic Holocene temperature record from the Agassiz ice cap and Greenland ice sheet evolution. *Proc. Natl Acad. Sci. USA* **114**, 5952–5957 (2017).
29. Walsh, J. E., Fetterer, F., Scott Stewart, J. & Chapman, W. L. A database for depicting Arctic sea ice variations back to 1850. *Geogr. Rev.* **107**, 89–107 (2017).
30. Kinnard, C. et al. Reconstructed changes in Arctic sea ice over the past 1,450 years. *Nature* **479**, 509–512 (2011).

Acknowledgements Funding was provided by US National Science Foundation (NSF) awards OPP-1205196 and PLR-1418256 to S.B.D., ARC-1205062 to B.E.S. and OPP-1205008 to M.J.E. L.D.T. acknowledges institutional support from Rowan University and the Doherty Postdoctoral Scholarship at Woods Hole Oceanographic Institution. M.B.O. acknowledges support from the Department of Defense Office of Naval Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a. Collection,

analysis and interpretation of core D5 was supported by NSF grant 0352511 to J.R.McC. B.P.Y.N. and M.R.v.d.B. acknowledge support from the Polar Program of the Netherlands Organization for Scientific Research (NWO/NPP) and the Netherlands Earth System Science Centre (NESSC). For running the MAR model, computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (FRS-FNRS) under grant number 2.5020.11 and the Tier-1 supercomputer (Zenobe) of the Fédération Wallonie Bruxelles infrastructure funded by the Walloon Region under grant agreement number 1117545. We thank M. Waszkiewicz and IDPO/IDDO for ice core drilling support. We thank the NSF Ice Core Facility (formerly NICL), A. York, M. Bingham, M. Hatch, S. Zarfos, Z. Li, and Milton Academy students for ice core sampling and processing support. We thank R. Banta for help with the D5 core, and A. Arienzo and N. Chellman for help in analysing the NU core. We thank M. Tedesco for providing the satellite melt duration data used in Fig. 3d. Maps in Figs. 1a, 3 and Extended Data Figs. 3, 4 were created with the NCAR Command Language (<https://www.ncl.ucar.edu>), and maps in Fig. 1b and Extended Data Fig. 6 were created with Esri ArcGIS. We acknowledge the use of Rapid Response imagery in Fig. 1b from the Land, Atmosphere Near real-time Capability for EOS (LANCE) system operated by the NASA/GSFC/Earth Science Data and Information System (ESDIS) with funding provided by NASA/HQ.

Reviewer information *Nature* thanks J. Briner and B. Vinther for their contribution to the peer review of this work.

Author contributions L.D.T. and S.B.D. conceived of and designed the study with input from M.B.O. B.E.S., S.B.D., M.J.E. and L.D.T. determined the ice core siting. S.B.D., L.D.T., M.B.O. and M.J.E. collected the ice cores. L.D.T. analysed melt stratigraphy for cores NU, GC and GW. S.B.D., J.R.McC. and L.D.T. analysed core D5. M.J.E. and J.R.McC. analysed ice core chemistry. Ice core chronology was led by M.B.O. with input from B.E.S., S.B.D., J.R.McC. and L.D.T. M.B.O. developed melt reconstruction code with input from L.D.T. B.P.Y.N. and M.R.v.d.B. provided RACMO2 model output and expertise. X.F. provided MAR model output and expertise. L.D.T. led the data analyses and interpretation, and wrote the manuscript, with input from S.B.D. and M.B.O. All authors read and commented on the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0752-4>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to L.D.T.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Ice-core melt analyses. Surface melt histories were developed from an array of percolation-zone firn and ice cores collected in CWG (Figs. 1, 2, Extended Data Table 1). Core chronology was achieved through annual layer counting of seasonal ice-core chemistry (including the summertime species $\delta^{18}\text{O}$, NO_3^- and SO_4^{2-} , and the winter/springtime species Mg^{2+} , Ca^{2+} and Na^+) for cores GC and GW. Soluble chemistry measurements were conducted on discrete, 5-cm-long longitudinal cuts for both cores using suppressed ion chromatography³¹. The relative dating uncertainty in the GW and GC cores is estimated at a year or less, based on validation of the annual layer counts with absolute-dated volcanic horizons³². In the NU core, simultaneous and depth-registered measurements of trace metallic species (Na, Mg, S, Cl, Ca, Br, Sr, Ce, Ti and Pb), black carbon, water isotopic abundance ($\delta^{18}\text{O}$ and δD of H_2O), as well as semiquantitative particle counts, were conducted at a continuous resolution of about 2 cm (water equivalent) following established methods^{33,34}. Annual layer counts were similarly determined in the NU core down to the Huaynaputina (Peru) volcanic horizon³² at AD 1601, via identification of winter–springtime seasonality in Na, Mg, Ca and particulates³⁵, and summertime seasonality in S, $\delta^{18}\text{O}$, black carbon and heavy metals³⁴. The relative uncertainty down to AD 1601 is estimated at two year or less using validation of annual layer counts with 15 absolute-dated tie points (13 volcanic, 2 radiogenic; following ref. ³⁶). Reference ³⁴ details the establishment of the depth–age scale for core D5.

All cores were latitudinally cut, cleaned and analysed for visible refrozen stratigraphic horizons at the US National Science Foundation Ice Core Facility, formerly the National Ice Core Laboratory (NICL). Cores GC, GW and NU were scanned with the NICL high-resolution (sub-millimetre) optical imaging system³⁷. Resulting images allowed for manual identification of refrozen melt layers using digitally registered core depths (see, for example, Extended Data Fig. 1a). Following established melt layer identification³⁸ and classification²⁶ methods, we characterized only horizontally continuous refrozen melt layers with discrete upper and lower bounds identifiable based on their optical contrast with surrounding undisturbed firn or ice, while excluding vertical percolation features or layers extending <50% of the total ice-core width. For core D5, we applied the same methodology to detailed stratigraphic logs produced from analysis of the ice core on a back-lit table at NICL. We report results of these analyses for the full length of the GC, GW and D5 records, and for the upper 86 m of the NU core (corresponding to the year AD 1650), beyond which refrozen melt horizons become ambiguous owing to layer thinning and reduced visible contrast between refrozen layers and surrounding ice. As annual accumulation and melt layers advect downward, thinning of both annual accumulation and melt layers occurs^{26,39}. It is therefore necessary to account for this thinning as a function of increasing depth, and following previous studies^{18,26}, we characterize melt as a percentage of annual snow accumulation after converting each to water equivalent lengths using density measurements along the core and assuming a density of ice (0.917 g cm^{-3}) for refrozen melt layers. After converting melt thicknesses to melt per cent, cores GC, GW and D5 were composited to form the CWG stack by averaging and standardizing (mean set to 0 and standard deviation set to 1) the records. To facilitate comparison, the NU melt per cent record was also standardized based on its full time series (1650–2013). Melt intensity increases were calculated before standardizing.

Broader spatial representation of ice-core melt histories. To assess the potential of our ice-core melt records to indicate wider ice-sheet melt variability, we calculated correlations across space between the annual CWG melt stack and gridded model and satellite melt datasets (Fig. 3). To quantify these relationships across space, we assessed Spearman rank order correlations owing to the nonparametric distributions of the datasets as a result of nonlinear melt evolution^{26,27} and tested significance (at the $P < 0.01$ level) using a two-tailed Student's t -test. Two sophisticated reanalysis-driven regional climate–snowpack models were used in our analysis, as well as satellite-derived estimates of annual surface melt duration. First, the most recent version of RACMO2, RACMO2.3p2, forced by the ECMWF ERA-40 and ERA-Interim reanalyses over 1958–2013 and statistically downscaled to 1 km spatial resolution from 11 km native outputs^{5,40}, was used to assess correlations with melt magnitude, refreezing, and runoff (Fig. 3a–c). As an assessment of the sensitivity of these correlations to model selection, we also calculated correlations against the MARv3.7 model forced by ERA-Interim over 1979–2013 and similarly downscaled to 1 km from 7.5 km native outputs⁶. Despite incorporating a unique snow model, results using MARv3.7 show broadly similar results to those from RACMO2.3p2, including wide-ranging significant correlations between the CWG ice-core composite and simulated melt and runoff (Extended Data Fig. 3). We also quantified spatial correlations with observed melting from satellite passive microwave radiometers at 25 km resolution^{4,41} over the time period of common overlap with full daily satellite observations (1988–2013), and note similarly high spatial correlations between satellite-observed summer melt duration and melt variability from the ice cores (Fig. 3d). Finally, we report correlations only where the gridded dataset indicated that the respective melt variable occurred in at least 50%

of the years of common overlap (that is, 28 years for the 1958–2013 RACMO2 and 13 years for the 1988–2013 satellite time series).

Runoff reconstruction. Ice core-derived runoff reconstructions were performed through multilinear regression of the CWG and NU melt records as predictor (independent) variables against RACMO2-modelled annual runoff as the predictand (dependent variable) over the period of common overlap between the two datasets (1958–2013). This procedure allows us to develop a transfer function by calibrating the ice-core melt records to the RACMO2-simulated runoff, and thus extend the runoff record back in time using only the ice-core melt records. We used a nonparametric Monte Carlo approach to generate 95% confidence intervals to the reconstructions. Namely, we developed 10,000 surrogate pseudo-random predictor/predictand pairings over the full calibration interval (1958–2013) using the frequency-domain method⁴², following established methods⁴³. To test reconstruction skill, we calculated the coefficient of determination (r^2), reduction of error (RE), and coefficient of efficiency (CE) statistics⁴⁴. Both RE and CE may vary from 1 to $-\infty$, where $\text{RE} > 0$ and $\text{CE} > 0$ indicate a predictive power above the simple mean estimate of RACMO2-simulated runoff during the calibration interval. Indeed, positive RE and CE values are commonly invoked to demonstrate skillful reconstructions^{45,46}. Any positive value of RE and CE, however, does not necessarily signify statistical significance provided the predictor/predictand series represent varying degrees of autocorrelation. Thus, to estimate whether a calibration statistic is significant for the ice-core and RACMO2 data, we additionally used 10,000 Monte Carlo iterations to develop empirical probability density functions for the reconstruction statistics (r^2 , RE, CE) tuned to the exact autocorrelative properties and length of the observed predictor/predictand series, and against which significance was tested⁴³. Next, to test the sensitivity of the reconstruction to calibration period, we performed stepwise cross-validation tests⁴⁵. This was done by initially defining the calibration interval as the most recent contiguous two-thirds of years in the period of common overlap between the ice-core and RACMO2 datasets (1977–2013), such that the validation interval initially represented the oldest one-third of available years (1958–1976). The cross-validation calibration interval was then stepped to one year older (1976–2012), while the verification interval was defined as all remaining years (1958–1975, 2013). This was progressively repeated until the calibration interval represented the oldest two-thirds of the period of common overlap (1958–1994) and the verification interval represented the youngest contiguous one-third of years (1995–2013). After each iteration, validation statistics (r^2 , RE, CE) were calculated and significance levels were determined using the Monte Carlo approach described above.

We performed this procedure of multilinear regression between ice-core melt records and RACMO2-modelled runoff, estimation of reconstruction confidence intervals, and iterative calibration period and statistic assessment to reconstruct runoff across each of the 19 ICESat elevation-defined GrIS surface drainage basins⁴⁷ and for the ice sheet as a whole (Extended Data Fig. 5). For this analysis, we first smoothed all data series using a 5-year centred moving average, as this allows for the possibility of meltwater percolation to deeper layers (potentially important for the higher-melt NU core) while also remaining consistent with the approximate integral timescale in melt and its climatic drivers across the GrIS. In each case, area-integrated RACMO2 runoff (that is, across a specific basin or entire ice sheet) was used as the predictand and the NU and CWG melt records as predictors.

For most individual basins and for GrIS-integrated runoff (the focus of our study), the majority of calibration intervals produced CE and RE values much greater than zero and significant validation statistics, indicating broad and skillful runoff reconstructions (Extended Data Fig. 6). The northernmost GrIS basins, in particular basins 1 and 8, overall had the fewest calibration/validation intervals where validation statistics (r^2 , RE, CE) were found to be significant, indicating comparatively low skill in reconstructing RACMO2 runoff in this region (Extended Data Fig. 6). However, our ice-core melt records and derived reconstruction of basin 8.2 runoff moderately agree with observed air temperatures from the nearby Pituffik site since 1948 (Extended Data Table 3; also see next section), and our reconstructed late twentieth-century increases in melt across northern Greenland (Extended Data Fig. 5) are consistent with recently rapid warming and melting reported in the Canadian Arctic^{28,48}. Robust reconstruction of GrIS-integrated runoff, despite a relatively poor fit with RACMO2 across northern basins is probably the result of little runoff originating from northern Greenland (about 8%–13% on average annually according to RACMO2 and MAR, respectively), and because GrIS-integrated runoff is dominated by runoff from the central-southwestern and southern basins (Extended Data Fig. 5). In these more southerly regions, our runoff reconstruction (Extended Data Figs. 5, 6), relationships between our ice core-derived melt and modelled and satellite-observed melt (Fig. 3, Extended Data Fig. 3), and relationships between point-scale melt processes and GrIS-integrated melt and runoff (Extended Data Fig. 4) are all particularly strong. Nevertheless, collection and integration of compatible ice-core melt records from regions currently under-represented in our reconstruction could

provide an opportunity to improve upon basin-level reconstructions of GrIS melt and runoff.

As an additional check on our runoff reconstruction, we also compared ice-core-reconstructed runoff calibrated using RACMO2 against two versions of the MAR model⁶ (v3.5.2 and v3.7), forced by ECMWF reanalyses and down-scaled to 5 km and 1 km resolution, respectively. We note strong agreement at the ice-sheet scale among the models and reconstruction (Fig. 4a). Similar methodological agreement is found across most individual drainage basins, although it is also clear there are biases among the magnitudes of runoff simulated by models in some basins that compensate when integrated over the full GrIS (Extended Data Fig. 5), a feature that has persisted since previous versions of the models⁴⁹.

Comparisons with observed air temperatures. As a further assessment of our ice-core melt time series and the derived runoff reconstruction, we quantified relationships between these variables and observed air temperatures across coastal Greenland (Extended Data Table 3). We used data^{50,51} compiled and archived by the Danish Meteorological Institute (DMI) for all sites with near-continuous summer (JJA) observations that predate the onset of the ERA-40 reanalysis (1958) used to drive RACMO2 and which extend to at least 2013, the most recent year in the ice core records. As in our runoff reconstruction, we use the CWG and NU melt time series as predictor variables in a multilinear regression against observed air temperatures after calculating 5-year moving averages of the time series. We find overwhelmingly significant (see next section) relationships between the temperature time series and the combined CWG and NU melt records, as well as between reconstructed runoff integrated across surface drainage basins adjacent to the location of air-temperature observations (Extended Data Table 3). We do note, however, that in general stronger peak correlations exist between the ice-core melt records and melt simulated directly on the ice sheet from RACMO2 (Extended Data Fig. 5), suggesting that coastal temperature observations (some of which are far removed from the ice sheet) may poorly or only partially represent climatic and surface energy balance conditions over the ice sheet. Owing to observational paucity on the ice sheet proper, especially before the last two decades¹⁹, ice-core-derived melt records that directly capture melt processes are therefore of particularly high value.

Quantifying significance of paired time series. Statistical significance of correlations among air temperatures, melt records, and runoff reconstructions was assessed with a two-tailed Student's *t*-test (Fig. 2 inset; Extended Data Tables 2, 3). For analyses between paired time series smoothed with 5-year moving averages, we accounted for autocorrelation in the smoothed data using two distinct methods. First, following ref.⁵², we calculated reduced effective sample sizes, n^* , and thus reduced effective degrees of freedom (d.o.f. = $n^* - 2$), given inherent lag-1 autocorrelation. As a secondary test of statistical significance given the presence of sample autocorrelation in the paired 5-year smoothed series, and following our reconstruction methods above, we also implemented a nonparametric Monte Carlo-based method⁴² to generate 10,000 pseudo-random series with the exact length and autocorrelative properties of each time series. We then quantified correlations using these surrogate datasets against the original second time series, and estimated significance by calculating the exceedance probability of generating a stronger correlation by chance alone.

Trend onset and emergence timing. We examined the timing of the onset of the most recent phase of increasing GrIS meltwater runoff, as well as the emergence of GrIS runoff beyond a pre-industrial baseline, and compared these timings to those of reconstructed Arctic temperatures (that is, 'Arctic2k' in ref.²¹) and summer (July–September) sea ice extent²⁹ (Fig. 4b; Extended Data Table 4). For these analyses, we used the SiZer (Significant ZERO crossings of derivatives) method⁵³ following recent studies^{21,45}, to calculate the median onset of sustained, significant ($P < 0.1$) trends from smoothed time series of the original datasets using a suite of Gaussian kernel filters from 15 years to 50 years. Uncertainty in trend change-point timing is expressed here as the median absolute deviation of the change points determined using each of the 36 synthetic data series (smoothed using the 15–50-year filters). Next, we tested whether GrIS runoff has surpassed natural variability, commonly defined as $+2\sigma$ beyond a pre-industrial period^{21,54}. As in our assessments of the magnitude of melt increases, we defined the pre-industrial period for the Arctic temperature and runoff reconstruction datasets as the eighteenth century (1700–1799), which has been identified as a period predating the onset of industrial-era Arctic warming²¹. Emergence is defined here as the median timing at which the 15–50-year smoothed time series surpass and remain above the pre-industrial $+2\sigma$ level. As with trend onset timing, we estimated uncertainty in emergence timing as the median absolute deviation in the emergence timings determined using the various filter widths. Timing and uncertainty range in trend onset and emergence timing for these and other data series are shown as the vertical lines and shaded regions in Fig. 4b, and expressed numerically in Extended Data Table 4.

We also evaluated the trend onset and emergence timing from an observationally derived sea-ice reconstruction by Walsh et al.²⁹ and compare this dataset

with a longer, proxy-derived summer sea ice reconstruction from Kinnard et al.³⁰. To facilitate comparison, we performed a minor adjustment (that is, bias correction) by matching the mean of the Kinnard time series to the mean of the Walsh dataset over 1850–1899, a period of common overlap and of little sea-ice change, by subtracting 0.22×10^6 km² from the full Kinnard dataset. Likewise, we note that the lower temporal resolution (40-yr low-pass filtered) of the Kinnard dataset precluded accurate assessment of change points and emergence from this longer dataset. Owing to the shorter length (1850–present) of the observationally derived Walsh sea-ice reconstruction, it was necessary to define the baseline period for this dataset as 1850–1899. Nevertheless, we have confidence in our estimated timing of summer sea-ice decline and its forced emergence given the clear similarity of the observationally based and proxy-based sea-ice datasets (Fig. 4b). Because the Kinnard data show relatively stable summer sea ice before 1850, but with a slight decrease during the baseline period used for the other datasets (1700–1799), we believe our calculated climate emergence from the Walsh dataset to be a conservative estimate (that is, early in time). Indeed, if we utilize the bias-adjusted summer sea-ice extent from Kinnard over the earlier baseline of 1700–1799 with the standard deviation of the annual-resolution Walsh data (over 1850–1899), the median emergence is moved forward in time to 1966 ± 3 years, still predating the median emergence of GrIS runoff by 10 yr.

Nonlinear melt–temperature relationship. We extracted profiles of melt and temperature at our coring sites and from an accumulation zone ($> 1,200$ m) elevational transect spaced every 100 m down the centreline of basin 7.1, containing our three cores forming the CWG stack (outline shown in Fig. 1a; also Extended Data Fig. 6). Because RACMO2.3p2 and MARv3.7 represent different time spans and incorporate unique snow models, we used both in our assessment of melt–temperature sensitivity (Fig. 4c). For RACMO2.3p2, simulated melt at the NU core site was determined to be implausibly low (26.5 millimetres of water equivalent per year over 1994–2013) despite a relatively high air temperature over this time period (-1.52°C), and so we did not utilize this model at this site. In comparison, MARv3.7 simulates much greater melt (898 millimetres of water equivalent per year) and higher air temperature (-0.92°C) averaged over 1994–2013. These values align with the expected melt–temperature relationship derived from basin 7.1, but also greatly exceed accumulation rate determined from the NU core (around 300 millimetres of water equivalent per year). Because our analysis of the ice core chemistry and resulting dating do not indicate missing years (that is, net annual ablation), we believe MARv3.7 simulates unrealistically high melt at this site. We speculate that these discrepancies are at least partially due to the small size of the NU ice cap (about $4\text{ km} \times 4\text{ km}$), its more maritime climate, and the relatively coarse grid scale of the native climate model simulations (11 km and 7.5 km for RACMO2 and MAR, respectively). Because our in situ observations reveal that the NU site experiences higher melt than our GrIS cores (and thus it more closely represents conditions along the margins of the GrIS where runoff is prevalent), we include surrogate melt and temperature points from RACMO2.3p2 and MARv3.7 for the NU core site. As a potential analogue for conditions at the NU core site, we plot in Fig. 4c values derived from 1,900 m along the centerline of basin 7.1, where simulated melt rates are consistent with those derived from our NU core. From these data, we find that an exponential function well represents the nonlinear relationship between surface melting and near-surface air temperature. This result is consistent with previous ice-core melt reconstructions²⁶ and model-based studies in Greenland⁵⁵ and Antarctica²⁷.

To assess melt conditions associated with runoff initiation, using both RACMO2.3p2 and MARv3.7 we isolated melt flux values for the three highest-elevation sampling sites along the basin 7.1 centreline with multiple years of runoff simulated at > 10 millimetres of water equivalent per year. From these runoff-associated melt fluxes, we calculated the mean melt associated with initial runoff and the standard deviation among the mean melt fluxes (shown as dashed runoff line and shaded area in Fig. 4c). This level of melt linked to initial onset of runoff in recent years is a function of more frequent or sustained high-intensity melt that leads to formation of thicker refrozen melt layers before favouring runoff over retention^{8,56,57}.

Analysis of periodicities in ice-core melt records. Semi-periodic oscillations are apparent in our primary ice-core-derived melt records (Fig. 2). To explore this variability more quantitatively, and to discern potential signatures of climate modes known to affect Greenland, we performed analyses to identify spectral signatures and coherence between our melt records and climate indicators including the North Atlantic Oscillation⁵⁸, the Atlantic Multidecadal Oscillation, the Greenland Blocking Index⁵⁹, and air temperatures. Extended Data Fig. 2 displays a subset of these analyses. The upper four plots show power spectral density using the multi-taper method⁶⁰ for our two main melt records and the summertime North Atlantic Oscillation and the Greenland Blocking Index, two atmospheric modes known to influence Greenland climate and melt^{14,23,59}. We find relatively high spectral power in the 13–16-year range common to NU and CWG. This periodicity appears consistent with the North Atlantic Oscillation, which shows

moderate evidence of spectral energy centred around a periodicity of about 13 years, although the summertime North Atlantic Oscillation is characteristically noisy (Extended Data Fig. 2d). We also note a periodicity of about 8 years in CWG corresponding to moderately high power in the Greenland Blocking Index, and also previously described for the winter North Atlantic Oscillation⁵⁹ but not apparent here in the summertime North Atlantic Oscillation. Finally, we find strong, high spectral power at 60-year periodicity in our coastal ice cap record, NU, consistent with the known 60–80-year periodicity of Atlantic Multidecadal Oscillation^{61,62}. Cross-wavelet analysis⁶³ reveals the greatest coherence and stationarity between the NU melt record and the southwest Greenland JJA air-temperature composite⁵¹ (Extended Data Fig. 2e). This analysis reveals high common power in the ~10–25-yr range and generally in phase relationships, with periods of melt partially leading temperature that probably reflect meltwater percolation in this high-melt core.

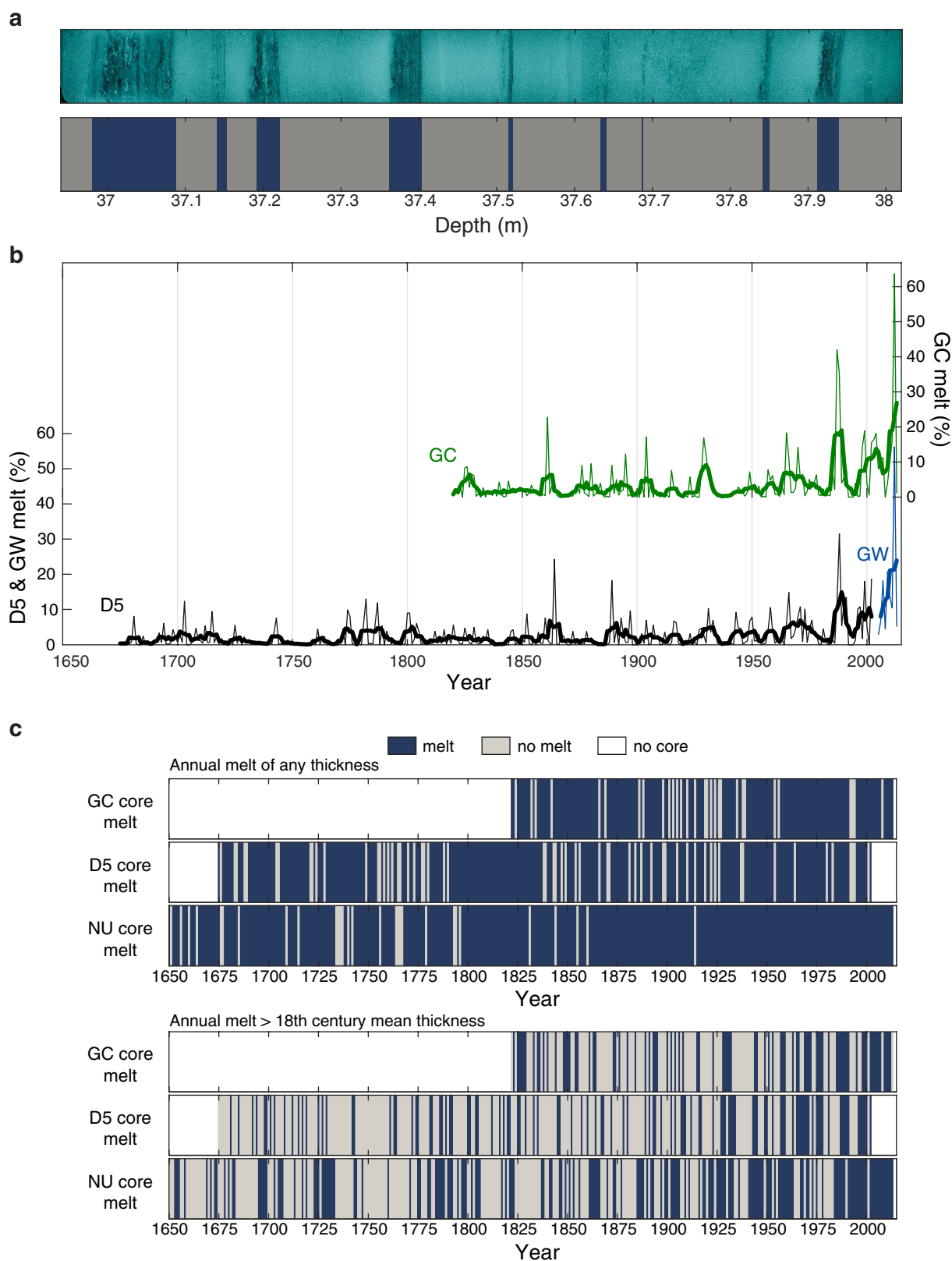
Our finding of strongest temporal coherence between NU and air temperature (as opposed to an individual climate mode) potentially indicates the combined influence of multiple climate modes on Greenland melt that are well represented by air temperature. Indeed, it is a confluence of processes that impact Greenland climate, combined with ice-sheet–climate interactions including the melt–albedo feedback, that govern the melt variability archived in ice cores. Given noise inherent to the relatively short time series, further analysis is warranted to better discern the varied factors responsible for the longer-term evolution of Greenland melt, and the representation of these processes in ice-core melt records.

Code availability. Code used for runoff reconstructions is available from L.D.T. upon request. Codes used for SiZer analysis were modified from ref. ²³ (<https://www.nature.com/articles/nature19082>) and available from L.D.T. upon request.

Data availability

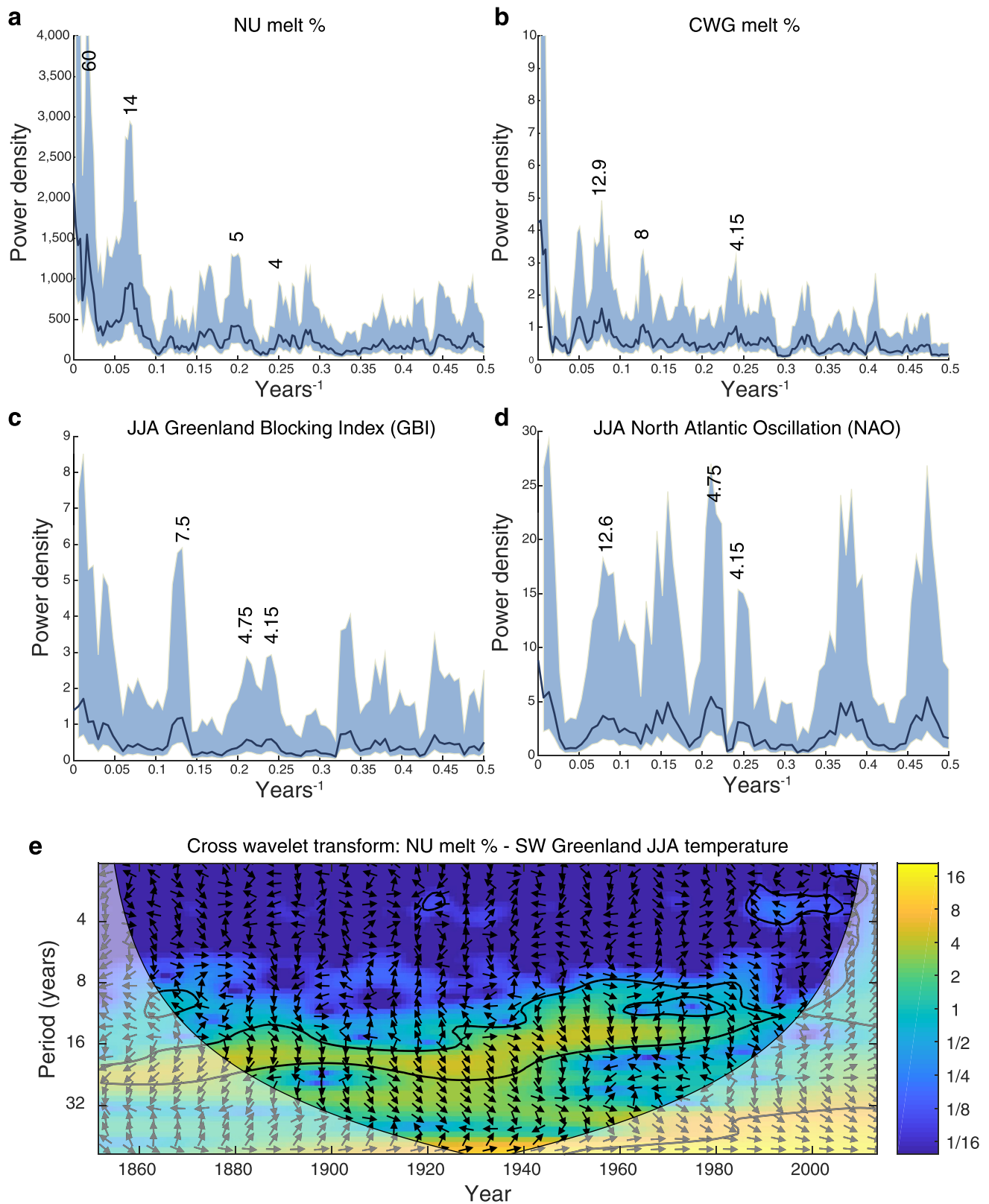
Ice-core melt records, the derived runoff reconstructions, and other records from cores NU, GC and GW are available via the NSF Arctic Data Center (<http://arcticdata.io>) and from the corresponding author upon request. Additionally, source data for Figs. 2, 4 are provided in the online version of this paper. RACMO2 model outputs⁵ as well as downscaled 1-km surface mass balance data are available from B.P.Y.N. and M.R.v.d.B. upon request. MAR model outputs⁶ are available from X.F. upon request. Greenland air-temperature data⁵¹ are available from <http://www.dmi.dk/laer-om/generelt/dmi-publikationer/tekniske-rapporter/>. Sea-ice data^{29,30} are available from <https://nsidc.org/data/g10010> and <https://www.nature.com/articles/nature10581>. Arctic air-temperature reconstruction data²¹ are available from <https://www.nature.com/articles/nature19082>. Satellite melt data^{4,41} are available from <http://www.cryocity.org>.

31. Curran, M. A. & Palmer, A. S. Suppressed ion chromatography methods for the routine determination of ultra low level anions and cations in ice cores. *J. Chromatogr. A* **919**, 107–113 (2001).
32. Sigl, M. et al. Timing and climate forcing of volcanic eruptions for the past 2,500 years. *Nature* **523**, 543–549 (2015).
33. McConnell, J. R., Lamorey, G. W., Lambert, S. W. & Taylor, K. C. Continuous ice-core chemical analyses using inductively coupled plasma mass spectrometry. *Environ. Sci. Technol.* **36**, 7–11 (2002).
34. McConnell, J. R. et al. 20th-century industrial black carbon emissions altered Arctic climate forcing. *Science* **317**, 1381–1384 (2007).
35. Gfeller, G. et al. Representativeness and seasonality of major ion records derived from NEEM firn cores. *Cryosphere* **8**, 1855–1870 (2014).
36. Arienzo, M. M. et al. A method for continuous ²³⁹Pu determinations in Arctic and Antarctic ice cores. *Environ. Sci. Technol.* **50**, 7066–7073 (2016).
37. McGwire, K. C. et al. An integrated system for optical imaging of ice cores. *Cold Reg. Sci. Technol.* **53**, 216–228 (2008).
38. Das, S. B. & Alley, R. B. Characterization and formation of melt layers in polar snow: observations and experiments from West Antarctica. *J. Glaciol.* **51**, 307–312 (2005).
39. Das, S. B. & Alley, R. B. Rise in frequency of surface melting at Siple Dome through the Holocene: evidence for increasing marine influence on the climate of West Antarctica. *J. Geophys. Res.* **113**, D02112 (2008).
40. Noël, B. et al. A daily, 1 km resolution data set of downscaled Greenland ice sheet surface mass balance (1958–2015). *Cryosphere* **10**, 2361–2377 (2016).
41. Tedesco, M. *Greenland Daily Surface Melt 25km EASE-Grid [1988–2013]* <http://www.cryocity.org/data.html> (City University of New York, New York, 2014).
42. Ebisuzaki, W. A method to estimate the statistical significance of a correlation when the data are serially correlated. *J. Clim.* **10**, 2147–2153 (1997).
43. Macias-Fauria, M., Grinsted, A., Helama, S. & Holopainen, J. Persistence matters: estimation of the statistical significance of paleoclimatic reconstruction statistics from autocorrelated time series. *Dendrochronologia* **30**, 179–187 (2012).
44. Cook, E. R., Briffa, K. R. & Jones, P. D. Spatial regression methods in dendroclimatology: a review and comparison of two techniques. *Int. J. Climatol.* **14**, 379–402 (1994).
45. Tierney, J. E. et al. Tropical sea surface temperatures for the past four centuries reconstructed from coral archives. *Paleoceanography* **30**, 2014PA002717 (2015).
46. Anchukaitis, K. J. et al. Last millennium Northern Hemisphere summer temperatures from tree rings. Part II, spatially resolved reconstructions. *Quat. Sci. Rev.* **163**, 1–22 (2017).
47. Zwally, H. J., Giovinetto, M. B., Beckley, M. A. & Saba, J. L. *Antarctic and Greenland Drainage Systems* http://icesat4.gsfc.nasa.gov/cryo_data/ant_gm_drainage_systems.php (GSFC Cryospheric Sciences Laboratory, NASA 2012).
48. Fisher, D. et al. Recent melt rates of Canadian arctic ice caps are the highest in four millennia. *Glob. Planet. Change* **84**, 3–7 (2012).
49. Vernon, C. L. et al. Surface mass balance model intercomparison for the Greenland ice sheet. *Cryosphere* **7**, 599–614 (2013).
50. Vinther, B. M., Andersen, K. K., Jones, P. D., Briffa, K. R. & Cappelen, J. Extending Greenland temperature records into the late eighteenth century. *J. Geophys. Res.* **111**, D11105 (2006).
51. Cappelen, J. (ed) *Greenland—DMI Historical Climate Data Collection 1784–2017 DMI Report 18-04* (DMI, Copenhagen, 2018).
52. Bretherton, C. S., Widmann, M., Dymnikov, V. P., Wallace, J. M. & Bladé, I. The effective number of spatial degrees of freedom of a time-varying field. *J. Clim.* **12**, 1990–2009 (1999).
53. Hannig, J. & Marron, J. S. Advanced distribution theory for SiZer. *J. Am. Stat. Assoc.* **101**, 484–499 (2006).
54. Hawkins, E. & Sutton, R. Time of emergence of climate signals. *Geophys. Res. Lett.* **39**, L01702 (2012).
55. Fettweis, X. et al. Estimating the Greenland ice sheet surface mass balance contribution to future sea level rise using the regional atmospheric climate model MAR. *Cryosphere* **7**, 469–489 (2013).
56. de la Peña, S. et al. Changes in the firn structure of the western Greenland Ice Sheet caused by recent warming. *Cryosphere* **9**, 1203–1211 (2015).
57. Noël, B. et al. A tipping point in refreezing accelerates mass loss of Greenland's glaciers and ice caps. *Nature Commun.* **8**, 14730 (2017).
58. Hurrell, J. & National Center for Atmospheric Research Staff (eds) *The Climate Data Guide: Hurrell North Atlantic Oscillation (NAO) Index (station-based)*. <https://climatedataguide.ucar.edu/climate-data/hurrell-north-atlantic-oscillation-nao-index-station-based> (NCAR, Boulder, 2003).
59. Hanna, E., Cropper, T. E., Hall, R. J. & Cappelen, J. Greenland Blocking Index 1851–2015: a regional climate change signal. *Int. J. Climatol.* **36**, 4847–4861 (2016).
60. Mann, M. E. & Lees, J. M. Robust estimation of background noise and signal detection in climatic time series. *Clim. Change* **33**, 409–445 (1996).
61. Schlesinger, M. E. & Ramankutty, N. An oscillation in the global climate system of period 65–70 years. *Nature* **367**, 723–726 (1994).
62. Trenberth, K. E. & Shea, D. J. Atlantic hurricanes and natural variability in 2005. *Geophys. Res. Lett.* **33**, L12704 (2006).
63. Grinsted, A., Moore, J. C. & Jevrejeva, S. Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Process. Geophys.* **11**, 561–566 (2004).



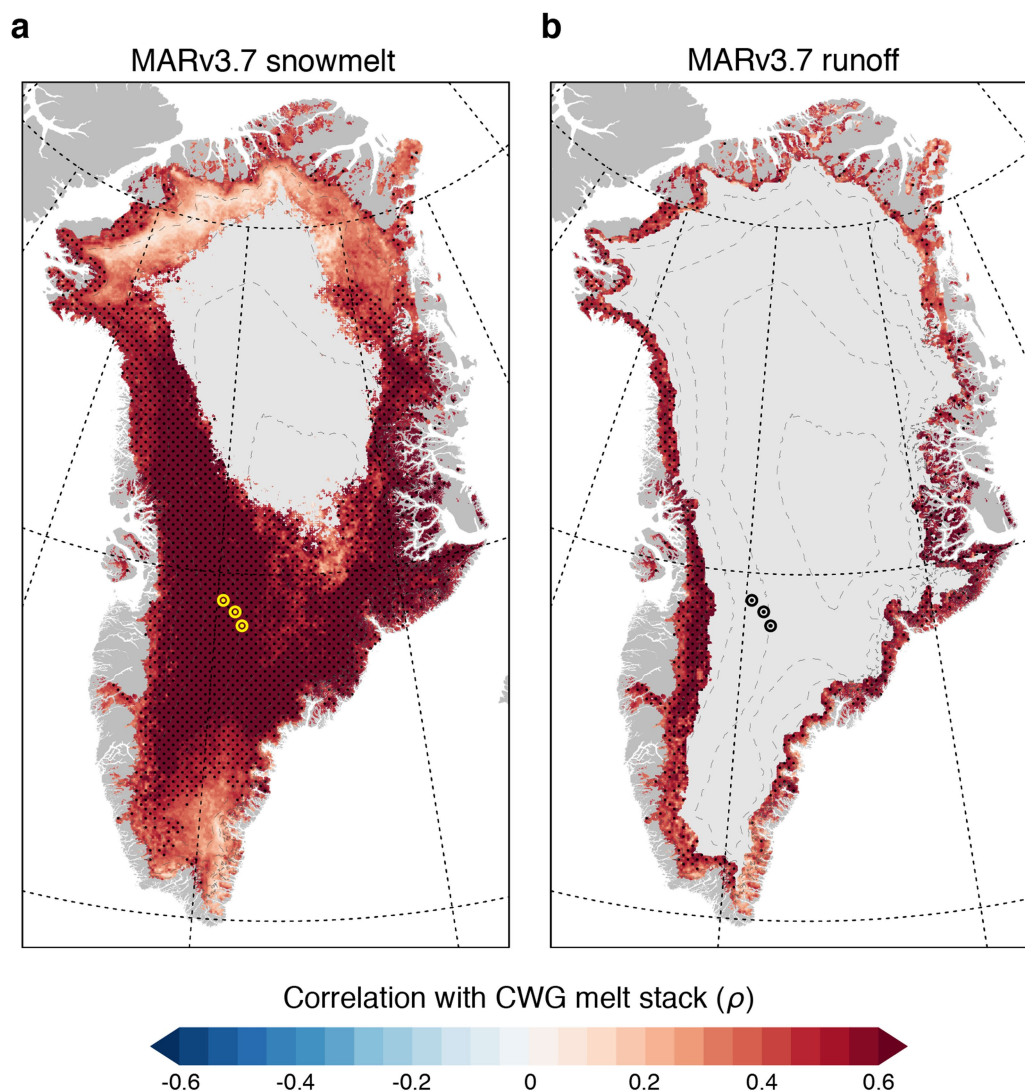
Extended Data Fig. 1 | Example core section and additional melt records. **a**, Example core scan image (top) and resulting digitized melt layers in blue (bottom). **b**, Annual melt per cent time series for cores in the CWG stack. Bold lines show 5-year moving averages. See Fig. 1 for locations. **c**, The top panel shows the presence (blue) or absence (grey) of any amount of refrozen surface melt within a particular year in our three longest ice cores, showing regular annual occurrence of melt at each

location. The bottom panel shows that, when filtered to show only years with melt percentages greater than the eighteenth-century mean melt at each site, a pattern towards recently more frequent, thicker (and thus more intense) melt emerges. As core GC does not span the eighteenth century, the mean eighteenth-century melt from the nearby D5 core was used as a baseline for GC as well.



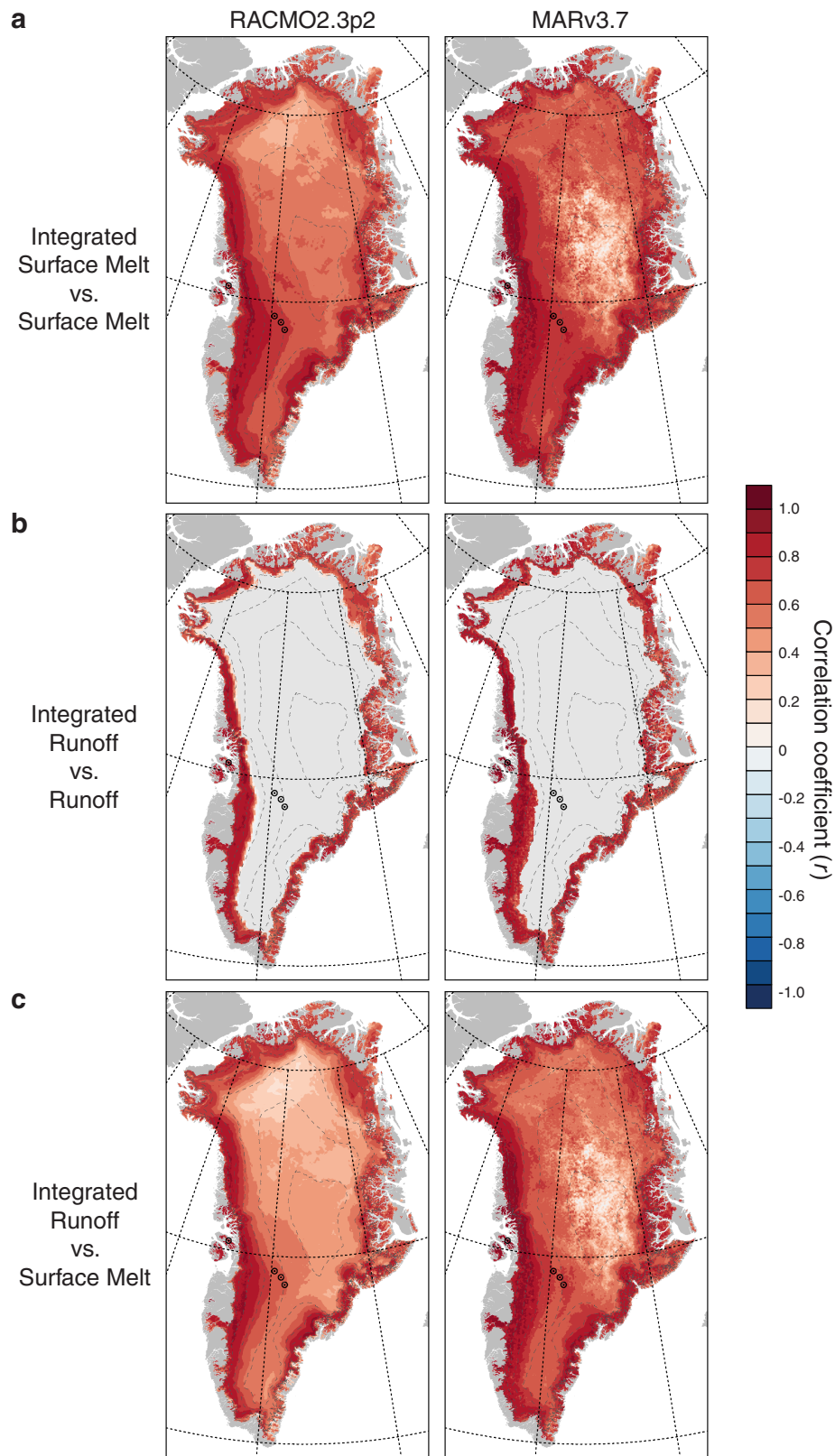
Extended Data Fig. 2 | Analysis of spectral signatures in melt and climate records. **a–d**, Multi-taper method spectral power plots for our ice cores, the Greenland Blocking Index⁵⁹ and the North Atlantic Oscillation⁵⁸. **e**, Cross-wavelet transform plot between NU melt and southwest Greenland temperatures⁵¹. In the spectral plots (**a–d**), years corresponding to specific peaks in spectral power are indicated by numbers, although many peaks are surrounded by a range of years with elevated spectral power. Shaded areas represent 95% confidence bounds, and we note that many peaks do not show up significantly (5% confidence) above a white-noise threshold. In the cross-wavelet plot in **e**, areas of

significant ($P < 0.05$) coherence are surrounded by a black line. The white-shaded areas represent regions where coherence cannot be confidently established owing to edge effects. Arrows indicate phase relationships: rightward (leftward) arrows indicate in-phase (out-of-phase) relationships, while downward (upward) arrows indicate melt leading (lagging) temperature. Our analyses found the strongest, and generally in-phase, coherence between NU and air temperature, as opposed to a single climate index, suggestive of the combined influence of multiple climate modes on GrIS melt that are well represented by air temperature (Methods).



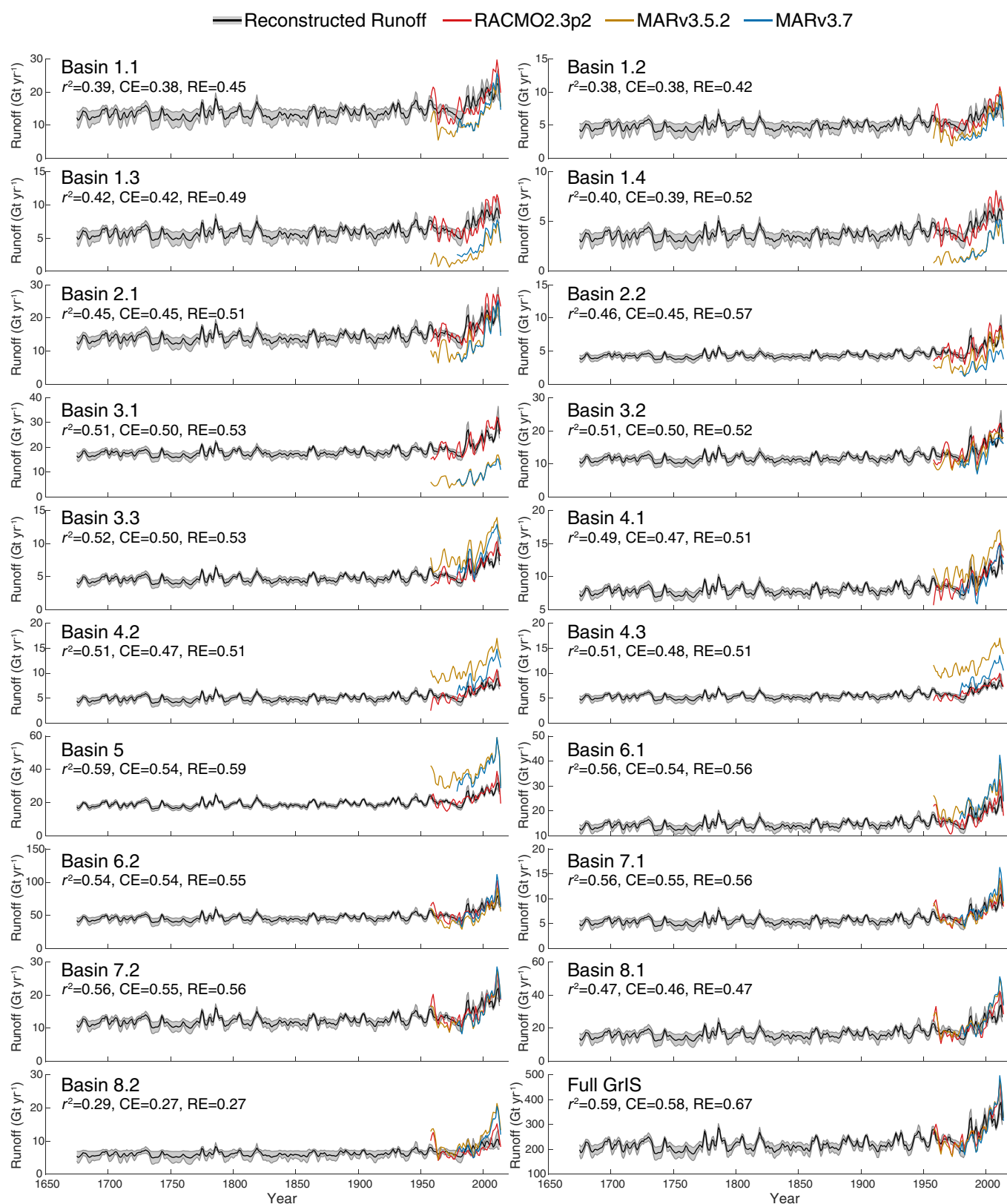
Extended Data Fig. 3 | Further evidence of spatially broad representation of melt processes in CWG cores. Spearman rank order correlations between the CWG melt stack and MARv3.7-modelled annual snowmelt (**a**) and snowmelt runoff (**b**) over the period 1978–2013. As in correlations against RACMO2 and satellite melt (Fig. 3), correlations

here are shown only for areas where MAR-simulated melt or runoff in at least 50% of the years of common overlap between the core and modelled datasets (18 years). Areas of significant correlation ($P < 0.01$) are denoted by a stipple pattern. Locations of cores used in the CWG stack are denoted by yellow (**a**) or black (**b**) points.



Extended Data Fig. 4 | Relationships between local and ice-sheet-integrated melt processes. Pearson correlation coefficients (r) between GrIS-integrated surface melt and surface melt in each grid cell (**a**), GrIS-integrated runoff and runoff in each grid cell (**b**), and GrIS-integrated runoff and surface melt in each grid cell (**c**). Correlations calculated for RACMO2.3p2 over 1958–2013 (left-hand plots) and

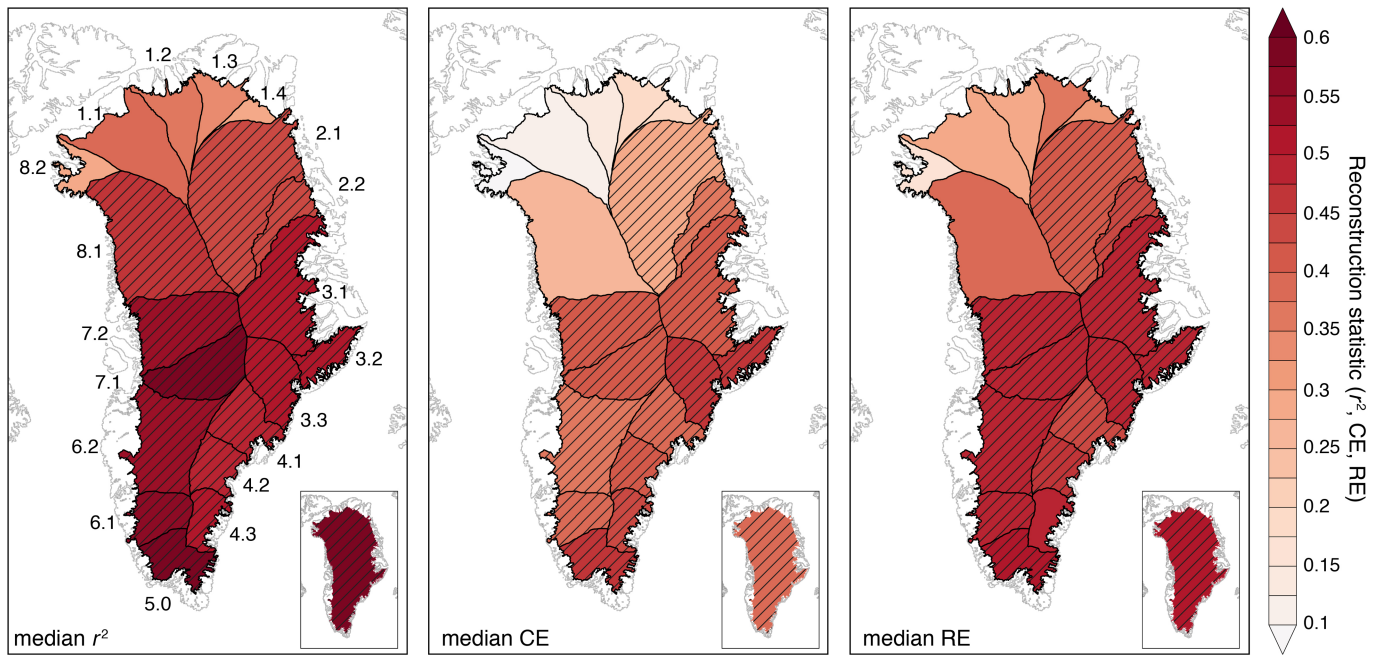
MARv3.7 over 1979–2013 (right-hand plots). Interannual variability in GrIS-integrated melt/runoff is well represented by local-scale interannual variability in melt/runoff. In situ melt records (for example, from well sited percolation-zone ice cores that are able to capture interannual melt variability) can therefore be used to quantify GrIS-integrated melt and runoff.



Extended Data Fig. 5 | Basin-specific runoff reconstructions.

Reconstructed runoff using the NU and CWG ice-core melt records calibrated to RACMO2.3p2 for 19 surface drainage basins⁴⁷ and for the full GrIS by summing each basin (lower right; as in Fig. 4). Note unique vertical axes. Reconstruction statistics shown are the maximum

values achieved for each metric across 20 stepwise calibration/validation intervals. Shaded regions around runoff reconstruction represent 95% confidence bounds. Modelled datasets smoothed using a 5-year Lowess filter. For details on reconstruction see Methods and see Extended Data Fig. 6 for basin location and further statistical assessment.



Extended Data Fig. 6 | Basin-specific and GrIS-integrated reconstruction statistics. Median values of runoff reconstruction skill statistics (r^2 , CE, RE) over 20 stepwise calibration/validation intervals, calculated for each surface drainage basin (basin numbers in left-hand plot), and for the GrIS as a whole (inset plots). Hatching denotes areas

where at least half of the calibration/validation intervals were found to be statistically significant at the $P < 0.1$ level, determined using 10,000 Monte Carlo simulations (see Methods). All basins had at least one calibration/validation interval where at least one of the three validation statistics was significant. See Methods for details on reconstruction methods.

Extended Data Table 1 | Details of firn and ice cores used in this study

Core	Year collected	Longitude	Latitude	Elevation (m a.s.l.)	Length (m)	Reference
GC	2015	-43.5025	68.892	2436	104	This study
GW	2014	-44.5198	69.2028	2291	9	This study
NU	2015	-52.2641	70.4865	1950	140	This study
D5	2003	-42.9	68.5	2515	148	McConnell et al., 2007 (ref. 34)

m.a.s.l., metres above sea level. The core source is from ref. ³⁴.

Extended Data Table 2 | Statistical relationships among ice cores and CWG air temperature

Correlation series	Annual/seasonal data	5-yr moving averages
CWG melt versus JJA air temperature ⁵⁰	$r = 0.331$ ($n = 170$)	$r = 0.581$ ($n^* = 95$)
NU melt versus JJA air temperature ⁵⁰	$r = 0.259$ ($n = 170$)	$r = 0.639$ ($n^* = 91$)
CWG melt versus NU melt	$r = 0.308$ ($n = 339$)	$r = 0.660$ ($n^* = 188$)

All correlations are significant at $P < 0.01$ level. For 5-year moving average time series, statistical significance was tested using reduced effective number of observations (n^*), as well as with a Monte Carlo-based procedure, to account for sample autocorrelation (Methods). Temperature data are from ref. ⁵⁰ as shown.

Extended Data Table 3 | Statistical relationships among ice cores, reconstructed runoff and pan-Greenland air temperatures

Observational record	Combined CWG and NU melt time series	Reconstructed runoff from adjacent basin(s)
	Multiple <i>r</i>	<i>r</i> / Runoff time series
Danmarkshavn (1949–2013)	0.839 (<i>n</i> = 65)	0.766 (<i>n</i> * = 39) / Basin 2.1
Ilulissat† (1850–2013)	0.665 (<i>n</i> = 174)	0.641 (<i>n</i> * = 96) / Basin 7.1
Ittoqqortoormiit (1949–2013)	0.767 (<i>n</i> = 65)	0.701 (<i>n</i> * = 36) / Basin 3.2
Ivituut / Narsarsuaq (1873–2013)	0.634 (<i>n</i> = 114)	0.599 (<i>n</i> * = 79) / Basin 5
Nuuk† (1866–2013)	0.543 (<i>n</i> = 148)	0.520 (<i>n</i> * = 82) / Basin 6.1
Pituffik (1948–2013)	0.545 (<i>n</i> = 66)	0.482 (<i>n</i> * = 37) / Basin 8.2
Tasiilaq (1895–2013)	0.348 (<i>n</i> = 119)	0.261§ (<i>n</i> * = 64) / Basin 4.2
Upernavik (1871–2013)	0.690 (<i>n</i> = 141)	0.643 (<i>n</i> * = 77) / Basin 8.1
Qaqortoq† (1873–2013)	0.463 (<i>n</i> = 141)	0.418 (<i>n</i> * = 79) / Basin 5
SW Greenland composite† (1840–2013)	0.595 (<i>n</i> = 174)	0.566 (<i>n</i> * = 95) / Basins 5, 6.1, 7.1

All series smoothed using a 5-year moving average to account for noise inherent to ice-core melt records. The second column shows correlations (multiple *r* values) using CWG and NU melt records as predictor variables against the indicated temperature time series, with all regressions significant at $P < 0.01$. The third column shows local correlations between our runoff reconstruction in the indicated basin(s) and the nearby temperature observations. For these series, we assessed significance using the reduced effective number of observations, *n**, accounting for serial correlation. All regressions are significant at $P < 0.01$ using *n** – 2 degrees of freedom, except for Tasiilaq‡. Relationships in bold were further found to be statistically significant assuming $P < 0.05$ using a Monte Carlo-based approach (see Methods for details).

†Includes infilled and regressed observations from regional stations^{50,51}.

‡Synthesis of Ilulissat, Nuuk and Qaqortoq master records^{50,51}.

§ $P = 0.037$ (using *n** – 2 degrees of freedom; see Methods).

Extended Data Table 4 | Timing of trend initiation and climate emergence

Climate record	Trend onset	Climate emergence
Arctic2k air temperatures (ref. 23)	1830.5 (1828–1833)	1937.5 (1936–1939)
Greenland runoff (full GrIS)	1865 (1857.5–1872.5)	1976 (1974–1978)
Arctic summer (JAS) sea ice (ref. 29)	1878.5 (1857.5–1899.5)	1947 (1943–1951)
Basin 1.1 reconstructed runoff	1865 (1858.5–1871.5)	1980.5 (1979–1982)
Basin 1.2 reconstructed runoff	1865 (1858.5–1871.5)	1983.5 (1983–1984)
Basin 1.3 reconstructed runoff	1865 (1858.5–1871.5)	1984 (1984–1984)
Basin 1.4 reconstructed runoff	1865 (1858.5–1871.5)	1984 (1983–1985)
Basin 2.1 reconstructed runoff	1865 (1857.5–1872.5)	1976 (1974–1978)
Basin 2.2 reconstructed runoff	1870 (1857–1883)	1969.5 (1966–1973)
Basin 3.1 reconstructed runoff	1868 (1858–1878)	1970.5 (1967–1974)
Basin 3.2 reconstructed runoff	1867 (1857.5–1867.5)	1971.5 (1968–1975)
Basin 3.3 reconstructed runoff	1866 (1858–1874)	1974 (1971.5–1976.5)
Basin 4.1 reconstructed runoff	1866 (1858–1874)	1975 (1972–1978)
Basin 4.2 reconstructed runoff	1865 (1858.5–1871.5)	1979 (1977–1981)
Basin 4.3 reconstructed runoff	1865 (1858.5–1871.5)	1981 (1980–1982)
Basin 5.0 reconstructed runoff	1865 (1858.5–1871.5)	1977.5 (1975–1980)
Basin 6.1 reconstructed runoff	1865 (1857.5–1872.5)	1976 (1974–1978)
Basin 6.2 reconstructed runoff	1865 (1857.5–1872.5)	1975.5 (1973–1978)
Basin 7.1 reconstructed runoff	1865 (1857.5–1872.5)	1976 (1973.5–1978.5)
Basin 7.2 reconstructed runoff	1865 (1857.5–1872.5)	1976 (1974–1978)
Basin 8.1 reconstructed runoff	1865 (1857.5–1872.5)	1977 (1975–1979)
Basin 8.2 reconstructed runoff	1865 (1858.5–1871.5)	1981 (1979–1983)

Onset of significant trends and climate emergence above pre-industrial levels is defined as 1700–1799 (median and range expressed as \pm median absolute deviation) for datasets shown in Fig. 4b (first three records) and for basin-level runoff reconstructions shown in Extended Data Fig. 5. Temperature and sea-ice data are from refs ^{23,29}.

The hippocampus is crucial for forming non-hippocampal long-term memory during sleep

Anuck Sawangjit¹, Carlos N. Oyanedel^{1,2}, Niels Niethard¹, Carolina Salazar¹, Jan Born^{1,3,4*} & Marion Inostroza^{1,4*}

There is a long-standing division in memory research between hippocampus-dependent memory and non-hippocampus-dependent memory, as only the latter can be acquired and retrieved in the absence of normal hippocampal function^{1,2}. Consolidation of hippocampus-dependent memory, in particular, is strongly supported by sleep^{3–5}. Here we show that the formation of long-term representations in a rat model of non-hippocampus-dependent memory depends not only on sleep but also on activation of a hippocampus-dependent mechanism during sleep. Rats encoded non-hippocampus-dependent (novel-object recognition^{6–8}) and hippocampus-dependent (object–place recognition) memories before a two-hour period of sleep or wakefulness. Memory was tested either immediately thereafter or remotely (after one or three weeks). Whereas object–place recognition memory was stronger for rats that had slept after encoding (rather than being awake) at both immediate and remote testing, novel-object recognition memory profited from sleep only three weeks after encoding, at which point it was preserved in rats that had slept after encoding but not in those that had been awake. Notably, inactivation of the hippocampus during post-encoding sleep by intrahippocampal injection of muscimol abolished the sleep-induced enhancement of remote novel-object recognition memory. By contrast, muscimol injection before remote retrieval or memory encoding had no effect on test performance, confirming that the encoding and retrieval of novel-object recognition memory are hippocampus-independent. Remote novel-object recognition memory was associated with spindle activity during post-encoding slow-wave sleep, consistent with the view that neuronal memory replay during slow-wave sleep contributes to long-term memory formation. Our results indicate that the hippocampus has an important role in long-term consolidation during sleep even for memories that have previously been considered hippocampus-independent.

Since the description of the patient H.M., who underwent bilateral removal of large portions of the hippocampus and suffered from severe anterograde amnesia, the distinction between hippocampus-dependent and non-hippocampus-dependent forms of memory has been widely accepted^{1,2,9}. Encoding and retrieval of hippocampus-dependent memories require the hippocampus, whereas this is not the case for non-hippocampus-dependent memory, which is otherwise comprised of rather heterogeneous kinds of memory (motor skills, cue fear conditioning and so on). The ‘standard consolidation theory’ and recent advances^{2,9,10} assume that memory of episodes, and in particular the relations among their elements, are initially encoded into hippocampal networks, but that during consolidation the representations are redistributed over days, weeks, and months to neocortical networks that serve as long-term stores. In this way these memories may become independent of the hippocampus^{9,11}.

Sleep is known to support memory consolidation^{3–5}. Sleep after memory encoding robustly enhances hippocampus-dependent memory, although there is also evidence that sleep enhances non-hippocampus-dependent forms of memory¹². With regard to

hippocampus-dependent memory, an active systems consolidation process has been proposed^{10,13,14} on the basis of findings that neural representations of freshly encoded memories are replayed during subsequent slow-wave sleep (SWS)^{15,16}. The neural replay originating from hippocampal networks, together with sharp-wave ripples and thalamic spindles, is likely to promote the transmission of memory information and, with repetitive occurrence, the gradual redistribution of the representation towards extrahippocampal networks^{17,18}.

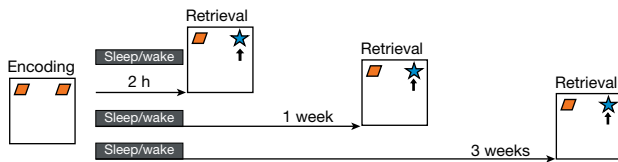
The ability of sleep to consolidate non-hippocampus-dependent memory is less well understood^{4,19,20}. Here, we compare the effects of post-encoding sleep with those of post-encoding wakefulness on consolidation of non-hippocampus-dependent and hippocampus-dependent forms of memory in rats, and examine the temporal evolution of consolidation effects. We used the novel-object recognition (NOR) task and an object–place recognition (OPR) task as tests of non-hippocampus-dependent and hippocampus-dependent memory, respectively (Fig. 1a). Performance on the NOR task relies on the perirhinal cortex, but normal hippocampal function is not necessary for encoding and retrieving NOR memory in rats^{6–8,21}.

After task encoding, rats either slept or remained awake during a 2-h interval. Retrieval was tested either immediately after the 2-h interval (recent test) or, in order to test long-term memory, 1 week or (for NOR only) 3 weeks later (remote tests; Fig. 1a). At the recent memory test, NOR memory did not differ between the sleep and wake conditions ($P = 0.43$), and exploration discrimination ratios indicated that there was significant NOR memory in both conditions ($P < 0.045$, Fig. 1b). By contrast, OPR memory at the recent test was enhanced in the sleep compared to the wake condition ($P = 0.034$) and was itself significant only after sleep ($P = 0.044$) and not in the wake condition ($P = 0.49$, $F_{1,20} = 4.70$, $P = 0.043$ for NOR/OPR \times sleep/wake analysis of variance (ANOVA) interaction). That sleep benefits recent OPR but not NOR memory confirms previous findings in rats^{22,23}, and has been taken as evidence that sleep preferentially strengthens hippocampus-dependent memory. Total object exploration, total distance travelled and mean speed at retrieval were comparable between sleep and wake conditions (all $P > 0.194$, Extended Data Fig. 1a), excluding confounds by nonspecific changes, for example, in locomotion or motivation.

At the remote test performed after 1 week, NOR memory still did not differ between the sleep and wake conditions ($P = 0.45$), and in both conditions rats showed significant NOR memory ($P < 0.045$). Also, as at the recent test, at the 1-week test rats showed better OPR memory in the sleep than the wake condition ($P = 0.001$), and OPR memory was not significant in the wake condition ($P > 0.308$, $F_{1,15} = 17.26$, $P = 0.001$ for NOR/OPR \times sleep/wake interaction; Fig. 1b).

NOR memory faded only when the retrieval delay was extended to 3 weeks. After 3 weeks, rats in the sleep condition but not in the wake condition showed significant NOR memory, and performance was significantly better for rats that had slept after encoding than for rats that had not ($P = 0.031$, $F_{1,17} = 4.696$, $P = 0.045$ for 1/3 weeks \times sleep/wake interaction in analysis of NOR data). A supplementary experiment indicated that the decrease in NOR memory after post-encoding

¹Institute of Medical Psychology and Behavioral Neurobiology, University of Tübingen, Tübingen, Germany. ²Graduate School of Neural and Behavioural Science, International Max Planck Research School, Tübingen, Germany. ³Center for Integrative Neuroscience, University of Tübingen, Tübingen, Germany. ⁴These authors jointly supervised this work: Jan Born, Marion Inostroza. *e-mail: jan.born@uni-tuebingen.de; marion.inostroza@uni-tuebingen.de

a Novel object recognition

Object place recognition

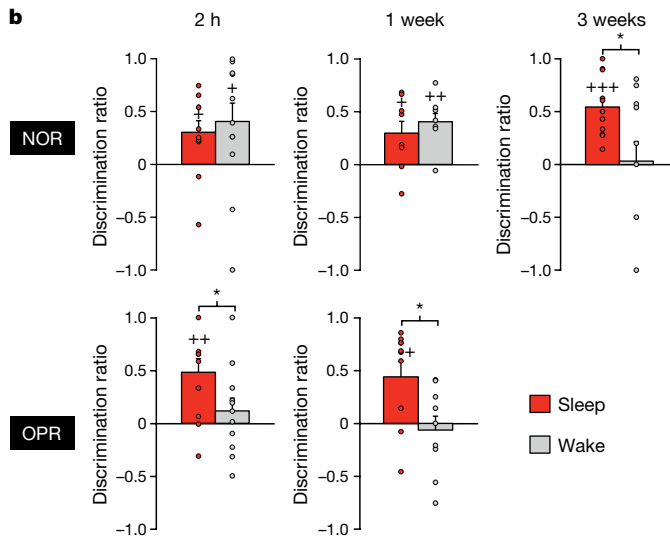
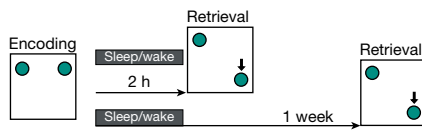


Fig. 1 | Effects of post-encoding sleep versus wakefulness on memory in the NOR and OPR tasks. **a**, During the encoding phase of both tasks, the rats explored (for 10 min) two identical objects in an arena. Encoding was followed by a 2-h interval in which the rat either slept or remained awake. Retrieval was tested immediately after the 2-h post-encoding interval (recent memory) and 1 week and (for NOR only) 3 weeks later (remote memory). At the retrieval test, the rat explored the arena for 5 min. To test NOR retrieval, one of the two objects (from the encoding phase) was replaced by a novel object (arrow); recognition memory was indicated when the rat spent more time exploring the novel object than the familiar object (discrimination ratio), with exploration during the first minute being most sensitive to exploration of novelty⁶. To test OPR retrieval, one of the objects was displaced (relative to its location at encoding, arrow) and memory for the place was indicated when the rat spent more time exploring the displaced object than the stationary object (which had not moved). **b**, Mean \pm s.e.m. discrimination ratios during the first minute of exploration for NOR and OPR at the recent (2 h) and remote (1 or 3 weeks) retrieval tests (dot plots overlaid). NOR memory benefited from post-encoding sleep (red bars; compared with wake, grey) only at the 3-week retrieval test, when NOR memory had decayed in the wake condition. By contrast, OPR memory benefited from sleep at both recent and remote testing. $n = 12, 8$ and 11 rats for NOR at 2 h, 1 week and 3 weeks; $n = 11$ and 9 rats for OPR at 2 h and 1 week, respectively. +++ $P < 0.001$, ++ $P < 0.01$, + $P < 0.05$ for one-sample t -tests against chance level; * $P < 0.05$ for pairwise t -tests (two-sided) between sleep and wake (see Extended Data Fig. 2 for discrimination ratios during the entire retrieval phase).

wakefulness had already occurred 1 week earlier, at a 2-week retrieval test (Extended Data Fig. 2c). Overall, remote testing confirmed that NOR memory was maintained over time periods of up to one week, even if encoding is followed by a wake period²⁴. However, the formation of more persistent long-term NOR memory requires sleep after encoding, with the sleep effect emerging only after 2–3 weeks, which corresponds to the time required for NOR memory in the wake condition to fade.

The consolidating effect of sleep on hippocampus-dependent spatial memory is mediated by repeated reactivations of the newly encoded hippocampal representations during subsequent SWS^{15,16,25}. Moreover, hippocampus-dependent and non-hippocampus-dependent memory systems have been found to interact during consolidation^{12,26}. Thus, we investigated whether hippocampal activity also critically contributes to the consolidation of non-hippocampus-dependent memory by reversibly inactivating hippocampal function by infusing muscimol into the dorsal hippocampus during sleep after encoding the NOR task.

At remote retrieval testing 3 weeks later, rats who had received muscimol injection into the hippocampus during sleep after learning did not show significant NOR memory ($P = 0.38$), whereas remote NOR memory was preserved in those injected with vehicle at the same time point ($P = 0.001$; $F_{1,14} = 8.99$, $P = 0.01$, for muscimol/vehicle main effect, Fig. 2a). Control parameters such as total object exploration did not differ between conditions, excluding nonspecific changes in motivation or vigilance (Extended Data Fig. 1). This result demonstrates that the hippocampus is crucial for the formation of persistent NOR memory during sleep. Previous studies that suppressed hippocampal activity after encoding in the NOR task had conflicting results^{8,27–29}, which fuelled a long-standing debate about the possible hippocampal dependency of NOR memory^{21,30,31}. These discrepancies can be resolved by our results, which show that formation of persistent long-term NOR memory relies on a hippocampal mechanism that is specifically active during sleep.

To determine whether the hippocampus is specifically involved in sleep consolidation, in a control experiment hippocampi were inactivated during a 2-h post-encoding wake period and retrieval was tested 1 week later. In these rats, NOR memory tended to be enhanced when compared to control animals whose hippocampal function was intact during the wake period after encoding ($F_{1,13} = 4.492$, $P = 0.054$ for muscimol/control main effect; Fig. 2b and Extended Data Fig. 3b), suggesting that, during wakefulness, hippocampal activity normally interferes with NOR memory consolidation⁸. Overall, these results corroborate the notion that persistent long-term NOR memory formation relies on a hippocampal mechanism that is specifically active during sleep, whereas non-hippocampal mechanisms during post-encoding wakefulness enable NOR memory over a period of 1 week.

We investigated whether the hippocampus would also be required for retrieval of long-term NOR memory at 3 weeks. Hippocampal infusion of muscimol before the 3-week retrieval test (in rats that had slept for 2 h after encoding) did not abolish NOR memory, with the rats' performance being closely comparable to that of a vehicle-infused group ($P > 0.70$ for all comparisons, Fig. 2a). This result indicates that whereas the formation of long-term NOR memory during sleep requires the hippocampus, its retrieval is not dependent on hippocampal function. In two further control experiments, muscimol was infused either shortly before a retrieval test that took place 30 min after encoding, or shortly before the encoding phase, with retrieval tested 30 min later (Fig. 2c). The experiments confirmed that short-term retrieval of NOR memory and encoding per se likewise do not depend on hippocampal function ($P = 0.46$ and $P = 0.79$, respectively, for differences between vehicle and muscimol)^{6,7}. Together, these results indicate that whereas the sleep-dependent formation of persistent long-term NOR memory requires the hippocampus, the retrieval of these memories is non-hippocampus-dependent at any time after encoding.

The architecture of post-encoding sleep was comparable to that reported in previous studies²³ (Extended Data Table 1). Correlation analyses revealed that remote NOR memory retrieval was strongly associated with measures of spindle activity during SWS, but not with rapid eye movement (REM)-sleep-related measures (Extended Data Table 2). Thus, NOR discrimination ratios at the 3-week retrieval test correlated with the number ($r = 0.719$, $P = 0.029$) and duration of spindles ($r = 0.705$, $P = 0.034$, Fig. 3a), with the latter correlation being most robust in an exploratory analysis focusing on the first 30 min of post-encoding sleep ($r = 0.888$, $P = 0.001$) in which neuronal replay in hippocampal networks, as a possible consolidation mechanism, is typically strongest³² (see Extended Data Fig. 4 for related OPR data).

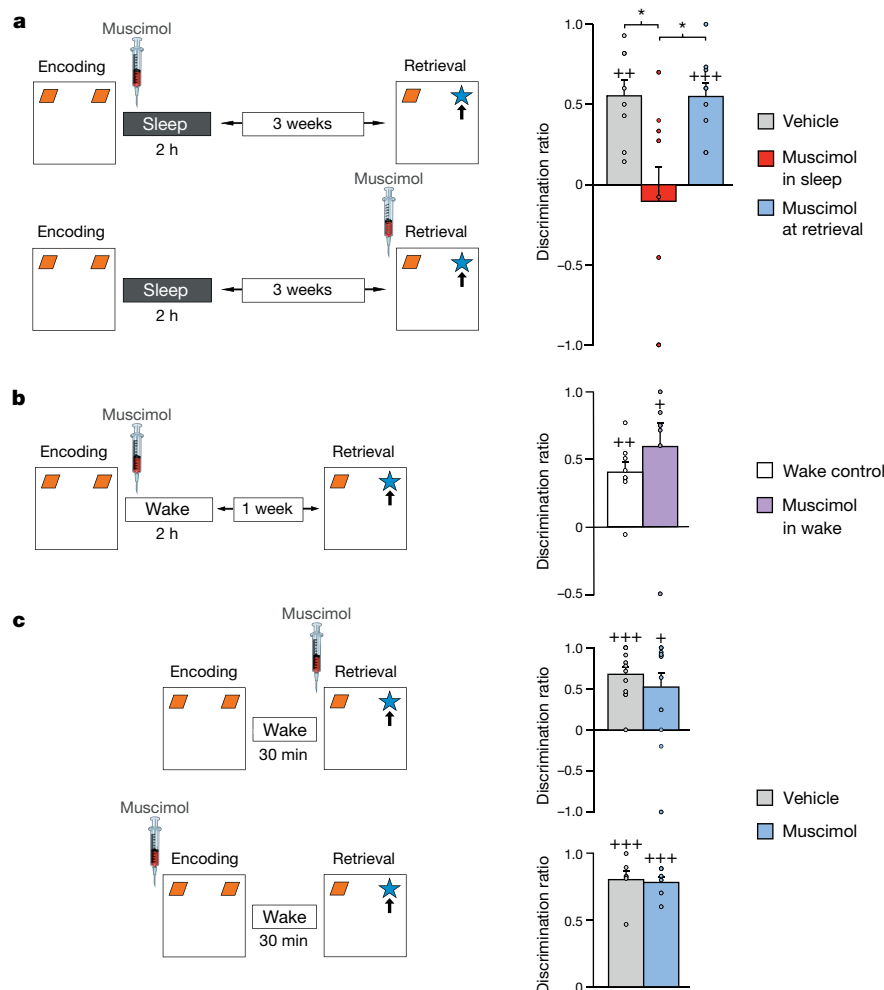


Fig. 2 | Effects of reversibly inactivating the hippocampus on NOR memory. Left, procedures; right, mean \pm s.e.m. discrimination ratios, with overlaid dot plots. **a**, To suppress hippocampal activity, muscimol was bilaterally infused (over 2 min) into the dorsal hippocampus, either during the post-encoding interval upon the first occurrence of continuous SWS (top; $n = 8$ rats each for muscimol and vehicle), or 15 min before remote retrieval testing 3 weeks after encoding (bottom; $n = 9$ rats). Hippocampal inactivation during post-encoding sleep (red bar) abolished remote NOR memory whereas inactivation before retrieval testing (blue bar) was ineffective. Grey bar, vehicle injection. **b**, Muscimol (purple bar) was infused shortly after encoding while the rats remained awake during the 2-h post-encoding interval ($n = 7$ rats). Retrieval was tested 1 week later. Compared with untreated wake control rats ($n = 8$ rats, empty bar),

which had intact hippocampal function and stayed awake during the post-encoding interval, hippocampal inactivation did not disturb but rather tended to enhance NOR performance. Timing (with reference to encoding), dosage and procedures of muscimol infusion were the same as in **a**. **c**, Muscimol (or vehicle) was infused 15 min before retrieval testing of recent NOR memory (top, $n = 12$ rats each) or 15 min before the encoding phase (bottom, $n = 6$ rats each). Retrieval was tested 30 min after encoding (rats stayed awake during this interval). Hippocampal inactivation does not affect retrieval of recent NOR memory either during retrieval or during encoding. $+++P < 0.001$, $++P < 0.01$, $+P < 0.05$ for one-sample t -test against chance level; $*P < 0.05$ for pairwise tests (two-sided) between conditions.

Intrahippocampal injection of muscimol during post-encoding sleep reduced electroencephalogram (EEG) theta activity ($P = 0.014$), which is thought to be generated in septal–hippocampal circuitry, and accordingly reduced time spent in both REM (0.83 ± 0.83 versus 6.04 ± 1.07 min after vehicle) and preREM sleep (1.97 ± 0.51 versus 5.80 ± 0.70 min after vehicle, both $P < 0.003$). Muscimol did not influence surface EEG activity during SWS (all $P > 0.410$, Extended Data Table 1). However, intrahippocampal local field potential (LFP) recordings from additional rats showed a distinct reduction in the number and density of hippocampal ripples, hippocampal spindle power and slow oscillation amplitude following muscimol infusion during post-encoding sleep ($P = 0.005$, 0.025 , 0.013 , and 0.007 , respectively; Fig. 3b). These changes are consistent with the view that muscimol prevents formation of long-term NOR memory by suppressing hippocampal ripples and associated reactivation of representations during SWS³³, although our findings do not rule out contributions of REM-sleep-related mechanisms³⁴.

There is ample evidence that the hippocampus is involved in the consolidation of memory classified as hippocampus-dependent, as

it can be acquired and retrieved only with normal hippocampal function⁹. We have now shown that normal hippocampal function is also required for the formation of persistent long-term representations on a task that, based on the same criterion, is classified as non-hippocampus-dependent^{6,7}. How does the hippocampus contribute to the formation of long-term NOR memory? In the NOR task, representation of the object resides mainly in the perirhinal cortex, whereas the hippocampus encodes spatial context features^{7,35}. Accordingly, retrieval in the NOR task also involves hippocampal function—making the task seemingly hippocampus-dependent—when it is performed in a context that is novel to the rat^{8,36}. Along this line, we propose that, during sleep, the hippocampus is likely to boost object representation through activation of context-related representations, rather than directly enhancing perirhinal object memory. The observed correlation of long-term NOR performance with post-encoding sleep spindle activity corroborates this view: neuronal reactivations of spatial context representations during sleep occur in the hippocampus, in conjunction with ripples and thalamic spindles^{25,32}. Spindles, moreover,

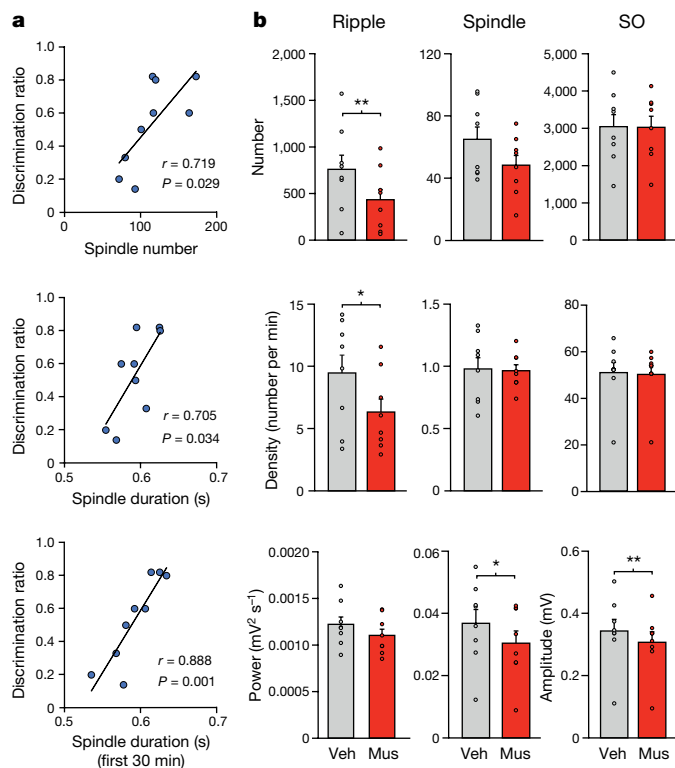


Fig. 3 | Contribution of post-encoding slow wave sleep to remote NOR memory. **a**, NOR performance (discrimination ratio) at the 3-week retrieval test was correlated with the number of sleep spindles during SWS (top) and spindle mean duration during the 2-h post-encoding interval (middle), as well as with spindle mean duration during the first 30 min of post-encoding sleep (bottom; Pearson's product-moment correlations, $n = 9$ rats). **b**, Intrahippocampal LFPs were recorded in additional rats to examine the effects of bilateral intrahippocampal infusion of muscimol (Mus) (versus vehicle, Veh) on (from left to right) ripples, spindles, and slow oscillations (SO) in hippocampal networks during SWS ($n = 8$ tests per condition). Muscimol decreased the total number and density of ripples, as well as spindle power and slow oscillation amplitude. Data shown as mean \pm s.e.m. with overlaid dot plots. $^{**}P < 0.01$, $^{*}P < 0.05$ for pairwise two-sided t -tests.

have been identified as a mechanism that favours the spreading of reactivations to extrahippocampal networks^{17,37}, thereby promoting plastic synaptic changes that can ultimately strengthen these extrahippocampal representations^{18,38}.

In conclusion, our findings suggest that a common hippocampal mechanism boosts consolidation in both hippocampus-dependent and non-hippocampus-dependent memory systems through the reactivation of contextual features. Indeed, in humans, hippocampal activity during training predicts sleep-dependent consolidation of a motor skill that is considered to be non-hippocampus-dependent^{12,26}. From this perspective, the formation of long-term representations during sleep, whether hippocampus-dependent or not, critically depends on their being encoded within a spatiotemporal context—that is, as episodic memories. Because such a mechanism puts the hippocampus-dependent episodic memory system into a supra-ordinate position to organize long-term memory, it has strong implications for current theorizing about interacting ‘parallel memory systems’³⁹. However, non-hippocampus-dependent memory is heterogeneous, and other memories of this kind need to be studied to scrutinize the proposed general hippocampal mechanism of long-term memory formation.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0716-8>.

Received: 6 April 2018; Accepted: 26 September 2018;
Published online 14 November 2018.

- Squire, L. R. Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychol. Rev.* **99**, 195–231 (1992).
- Moscovitch, M., Cabeza, R., Winocur, G. & Nadel, L. Episodic memory and beyond: the hippocampus and neocortex in transformation. *Annu. Rev. Psychol.* **67**, 105–134 (2016).
- Stickgold, R. Sleep-dependent memory consolidation. *Nature* **437**, 1272–1278 (2005).
- Rasch, B. & Born, J. About sleep's role in memory. *Physiol. Rev.* **93**, 681–766 (2013).
- Tononi, G. & Cirelli, C. Sleep and the price of plasticity: from synaptic and cellular homeostasis to memory consolidation and integration. *Neuron* **81**, 12–34 (2014).
- Winters, B. D., Forwood, S. E., Cowell, R. A., Saksida, L. M. & Bussey, T. J. Double dissociation between the effects of peri-posterior cortex and hippocampal lesions on tests of object recognition and spatial memory: heterogeneity of function within the temporal lobe. *J. Neurosci.* **24**, 5901–5908 (2004).
- Winters, B. D., Saksida, L. M. & Bussey, T. J. Object recognition memory: neurobiological mechanisms of encoding, consolidation and retrieval. *Neurosci. Biobehav. Rev.* **32**, 1055–1070 (2008).
- Oliveira, A. M., Hawk, J. D., Abel, T. & Havekes, R. Post-training reversible inactivation of the hippocampus enhances novel object recognition memory. *Learn. Mem.* **17**, 155–160 (2010).
- Eichenbaum, H. A cortical-hippocampal system for declarative memory. *Nat. Rev. Neurosci.* **1**, 41–50 (2000).
- McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* **102**, 419–457 (1995).
- Frankland, P. W. & Bontempi, B. The organization of recent and remote memories. *Nat. Rev. Neurosci.* **6**, 119–130 (2005).
- King, B. R., Hoedlmoser, K., Hirschauer, F., Dolfin, N. & Albouy, G. Sleeping on the motor engram: the multifaceted nature of sleep-related motor memory consolidation. *Neurosci. Biobehav. Rev.* **80**, 1–22 (2017).
- Diekelmann, S. & Born, J. The memory function of sleep. *Nat. Rev. Neurosci.* **11**, 114–126 (2010).
- Lewis, P. A. & Durrant, S. J. Overlapping memory replay during sleep builds cognitive schemata. *Trends Cogn. Sci.* **15**, 343–351 (2011).
- Wilson, M. A. & McNaughton, B. L. Reactivation of hippocampal ensemble memories during sleep. *Science* **265**, 676–679 (1994).
- Rasch, B., Büchel, C., Gais, S. & Born, J. Odor cues during slow-wave sleep prompt declarative memory consolidation. *Science* **315**, 1426–1429 (2007).
- Latchoumane, C. V., Ngo, H. V., Born, J. & Shin, H. S. Thalamic spindles promote memory formation during sleep through triple phase-locking of cortical, thalamic, and hippocampal rhythms. *Neuron* **95**, 424–435.e6 (2017).
- Seibt, J. et al. Cortical dendritic activity correlates with spindle-rich oscillations during sleep in rodents. *Nat. Commun.* **8**, 684 (2017).
- Ramanathan, D. S., Gulati, T. & Ganguly, K. Sleep-dependent reactivation of ensembles in motor cortex promotes skill consolidation. *PLoS Biol.* **13**, e1002263 (2015).
- Li, W., Ma, L., Yang, G. & Gan, W. B. REM sleep selectively prunes and maintains new synapses in development and learning. *Nat. Neurosci.* **20**, 427–437 (2017).
- Brown, M. W. & Aggleton, J. P. Recognition memory: what are the roles of the perirhinal cortex and hippocampus? *Nat. Rev. Neurosci.* **2**, 51–61 (2001).
- Inostroza, M., Binder, S. & Born, J. Sleep-dependency of episodic-like memory consolidation in rats. *Behav. Brain Res.* **237**, 15–22 (2013).
- Oyanedel, C. N. et al. Role of slow oscillatory activity and slow wave sleep in consolidation of episodic-like memory in rats. *Behav. Brain Res.* **275**, 126–130 (2014).
- Broadbent, N. J., Gaskin, S., Squire, L. R. & Clark, R. E. Object recognition memory and the rodent hippocampus. *Learn. Mem.* **17**, 5–11 (2009).
- van de Ven, G. M., Trouche, S., McNamara, C. G., Allen, K. & Dupret, D. Hippocampal offline reactivation consolidates recently formed cell assembly patterns during sharp wave-ripples. *Neuron* **92**, 968–974 (2016).
- Albouy, G. et al. Maintaining vs. enhancing motor sequence memories: respective roles of striatal and hippocampal systems. *Neuroimage* **108**, 423–434 (2015).
- de Lima, M. N., Luft, T., Roesler, R. & Schröder, N. Temporary inactivation reveals an essential role of the dorsal hippocampus in consolidation of object recognition memory. *Neurosci. Lett.* **405**, 142–146 (2006).
- Rosato, J. I. et al. On the role of hippocampal protein synthesis in the consolidation and reconsolidation of object recognition memory. *Learn. Mem.* **14**, 36–46 (2007).
- Kim, J. M., Kim, D. H., Lee, Y., Park, S. J. & Ryu, J. H. Distinct roles of the hippocampus and perirhinal cortex in GABA_A receptor blockade-induced enhancement of object recognition memory. *Brain Res.* **1552**, 17–25 (2014).
- Cohen, S. J. & Stackman, R. W. Jr Assessing rodent hippocampal involvement in the novel object recognition task. A review. *Behav. Brain Res.* **285**, 105–117 (2015).
- Squire, L. R., Wixted, J. T. & Clark, R. E. Recognition memory and the medial temporal lobe: a new perspective. *Nat. Rev. Neurosci.* **8**, 872–883 (2007).
- O'Neill, J., Pleydell-Bouverie, B., Dupret, D. & Csicsvari, J. Play it again: reactivation of waking experience and memory. *Trends Neurosci.* **33**, 220–229 (2010).

33. Girardeau, G., Benchenane, K., Wiener, S. I., Buzsáki, G. & Zugaro, M. B. Selective suppression of hippocampal ripples impairs spatial memory. *Nat. Neurosci.* **12**, 1222–1223 (2009).
34. Chen, L., Tian, S. & Ke, J. Rapid eye movement sleep deprivation disrupts consolidation but not reconsolidation of novel object recognition memory in rats. *Neurosci. Lett.* **563**, 12–16 (2014).
35. Brown, M. W., Barker, G. R., Aggleton, J. P. & Warburton, E. C. What pharmacological interventions indicate concerning the role of the perirhinal cortex in recognition memory. *Neuropsychologia* **50**, 3122–3140 (2012).
36. Piterkin, P., Cole, E., Cossette, M. P., Gaskin, S. & Mumby, D. G. A limited role for the hippocampus in the modulation of novel-object preference by contextual cues. *Learn. Mem.* **15**, 785–791 (2008).
37. Bergmann, T. O., Mölle, M., Diedrichs, J., Born, J. & Siebner, H. R. Sleep spindle-related reactivation of category-specific cortical regions after learning face-scene associations. *Neuroimage* **59**, 2733–2742 (2012).
38. Chauvette, S., Seigne, J. & Timofeev, I. Sleep oscillations in the thalamocortical system induce long-term neuronal plasticity. *Neuron* **75**, 1105–1113 (2012).
39. Lisman, J. et al. Viewpoints: how the hippocampus contributes to memory, navigation and cognition. *Nat. Neurosci.* **20**, 1434–1447 (2017).
40. Miller, K. J., Botvinick, M. M. & Brody, C. D. Dorsal hippocampus contributes to model-based planning. *Nat. Neurosci.* **20**, 1269–1276 (2017).
41. Pack, A. I. et al. Novel method for high-throughput phenotyping of sleep in mice. *Physiol. Genomics* **28**, 232–238 (2007).
42. Neckelmann, D., Olsen, O. E., Fagerland, S. & Ursin, R. The reliability and functional validity of visual and semiautomatic sleep/wake scoring in the Møll-Wistar rat. *Sleep* **17**, 120–131 (1994).

Acknowledgements We thank I. Sauter for technical support and E. Coffey and E. Bolinger for proof reading. This study was supported by a grant from the Deutsche Forschungsgemeinschaft (Tr-SFB 654). A.S. received a scholarship from the Development and Promotion of Science and Technology Talented Project (DPST), Thailand.

Reviewer information Nature thanks J. Csicsvari, S. Ramirez and R. Stickgold for their contribution to the peer review of this work.

Author contributions A.S., J.B. and M.I. planned and designed the experiments and wrote the manuscript. A.S. and C.S. performed the experiments and the histology. A.S., C.N.O. and N.N. analysed the data. All authors approved the final version of the paper.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0716-8>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0716-8>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to J.B. or M.I.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Animals. Ninety-one adult male Long Evans rats (Janvier, 260–310 g, 10–12 weeks) were used for the experiments. Rats were housed in groups of 2–4 rats per cage, except during the post-surgery recovery period, when they were kept individually on a 12-h light/12-h dark cycle (lights on at 06:00), and had unrestricted access to water and food throughout the experiments. All experimental procedures were performed in accordance with the European animal protection laws and policies (Directive 86/609, 1986, European Community) and were approved by the Baden-Württemberg state authority.

No statistical methods were used to predetermine sample size. In all experiments, rats were randomly assigned to experimental groups and conditions before the experiment. The experimenters were not blinded to the experimental group or condition during data collection. However, all behavioural and electrophysiological recordings were analysed offline, with the experimenters blinded to the experimental groups and conditions.

Design and general procedures. Different groups of rats were tested on either the NOR task or the OPR task, using post-encoding retention intervals of 2 h, 1 week and, only for the NOR task, of 3 weeks. Each group of animals was tested on a sleep condition (allowed to sleep during the 2-h post-encoding interval) and a wake condition (stayed awake during this interval). The order of sleep and wake conditions was counterbalanced across animals of a group. For an individual rat, the conditions were separated by an interval that was at least 2 weeks and twice as long as the tested retention interval. Encoding and the subsequent 2-h post-encoding interval took place in both the sleep and wake conditions during the animal's rest phase (between 08:00 and 13:00). In the sleep condition, during the 2-h post-encoding interval, the animals were left undisturbed in a 'post-encoding' box (35 × 35 cm, height: 45 cm) that was made of plastic and contained some bedding materials. Sleep was assessed using video recorded behaviour using standard procedures (see below). In the wake condition, wakefulness was enforced using gentle handling^{22,23}. This procedure minimizes stress and confounding influences of locomotion. It involved tapping on the retention box and, if necessary, gently shaking the box. No intense stimulation was used, and video records ensured that signs of startle or freezing behaviour did not occur. In the remote groups tested after 1 and 3 weeks, animals were brought to their home cages after the 2-h post-encoding interval and kept under routine conditions until testing.

Habituation and memory tasks. After handling daily for five consecutive days for 5–10 min, the rats were brought into the test room once every day on three consecutive days for a habituation session. For object familiarization, the rat was placed into an empty cage with an object (not used for the experiments) positioned in the centre of the cage. The rat was allowed to freely explore the object for 10 min. For arena familiarization, the rat was then placed into an empty open field, facing a different wall of the open field at each session to facilitate allocentric navigation, and allowed to explore for 10 min. Immediately afterwards, the rat was left undisturbed in the post-encoding box for 2 h.

On the day after the habituation phase, the experiment started with the encoding phase of the memory task. The encoding phase was identical for the NOR and OPR task, and comprised a 10-min interval during which the rats were allowed to explore two identical objects in the open field. For testing retrieval on the NOR task, one of the two objects of the encoding phase was replaced by a novel object. For testing retrieval on the OPR task, one of the two objects of the encoding phase was moved to a different location. At each test, the rat had 5 min to explore the arena.

The tasks were performed in a room with a noise-generator providing masking noise. The open field (80 cm × 80 cm, height of walls: 40 cm) was made of grey PVC. Through the open upper side of the arena the rat could perceive distal cues (two rectangles at the north wall, two other rectangles at the east wall, and a square at the west wall). Objects for exploration were made of glass, with different colours and shapes, and heavy enough not to be moved by the rat (height: 15–30 cm; base diameter: 7–12 cm). They were positioned at least 10 cm equidistant from the walls to ensure that the animal's preference to stay in corners did not bias exploration times. Pilot studies ensured that the rats could discriminate among the different objects and did not show any preference for one of the objects. The locations of objects during the encoding and retrieval phases were randomized across rats. Each rat's exploration behaviour was monitored by a video camera and analysed offline by an experienced researcher using ANY-maze software (Stoelting Europe). After each phase, the apparatus and objects were cleaned with water containing 70% ethanol.

Inactivating the hippocampus during sleep. To reversibly inactivate the dorsal hippocampus during sleep, we infused the GABA-A receptor agonist muscimol, according to standardized procedures^{8,40}. After 5 days of handling, guide cannulae were surgically implanted bilaterally into the dorsal hippocampi, and at least 8 days were allowed for recovery. Muscimol (Sigma, 0.5 µg dissolved in 0.5 µl saline solution, per hemisphere) or an equivalent volume of vehicle (saline solution) was infused bilaterally over 2 min by an automated syringe pump. (In pilot studies with

this dosing, no spread of the substance to extrahippocampal regions occurred; Extended Data Fig. 5b.) For substance administration, two 30-gauge injection cannulae were connected to two 10-µl Hamilton microsyringes (Hamilton), with 1-m polyethylene cannula tubing. The injection cannulae protruded 1 mm beyond the tip of the guide cannulae. The injection cannulae were kept in the bilateral guide cannulae for a further 2 min to prevent backflow. The procedure enabled substance administration into freely moving rats without disturbing ongoing sleep. Rats were killed at the end of the experiments for histological confirmation of the infusion sites (Extended Data Fig. 5).

The effects of muscimol and vehicle were compared in a between-subjects comparison in 16 rats (8 per group). To test the effects of hippocampal inactivation during sleep in the 2-h post-encoding interval, substance administration started immediately upon (visual) online detection of continuous SWS for at least 10 s. On average, substance administration took place after 38.30 ± 2.16 min of the post-encoding interval in the muscimol condition and after 40.35 ± 1.33 min in the vehicle condition ($P = 0.42$).

Surgery in experiments with reversible inactivation of the hippocampus. Guide cannulae were implanted under general isoflurane anaesthesia (induction: 1–2%, maintenance: 0.8–1.2% in 0.35 l/min O₂). Preoperatively, fentanyl (0.005 mg/kg), midazolam (2 mg/kg) and medetomidine (0.15 mg/kg) were administered intraperitoneally. Rats were placed in the stereotaxic frame and the skull was exposed. Two stainless steel guide cannulae (7 mm long, 23 gauge, Plastics One) were bilaterally implanted into the dorsal hippocampi (anterior–posterior (AP): −4.3 mm, mediolateral (ML): ±2.8 mm, dorsoventral (DV): −1.3 mm under skull surface, relative to bregma). The cannulae were introduced to this position laterally tilted by 9° with respect to the vertical axis and were affixed to the skull with four bone screws and cold polymerizing dental resin. Dummy cannulae (7 mm long, Plastics One) were inserted into the guide cannulae and removed only for infusions.

For simultaneous EEG recordings in the animals, four screw electrodes were implanted: two frontal electrodes (AP: +2.6 mm, ML: ±1.8 mm, relative to bregma) and two occipital electrodes (AP: −10.0 mm, ML: ±1.8 mm), with the latter serving (for all recordings) as reference and ground, respectively. Additionally, in a subgroup of animals, two platinum electrodes were attached to the guide cannulae to record hippocampal LFP signals (AP: −4.3 mm, ML: ±2.8 mm, DV: −2.3 mm, relative to bregma). Two stainless steel wire electrodes were implanted bilaterally in the neck muscles for electromyography (EMG) recordings. Electrodes were connected to a Mill-Max pedestal and fixed to the skull with cold polymerizing dental resin and the wound was sutured. After the surgery, the rats received a subcutaneous 1-ml injection of saline solution to prevent dehydration, and carprofen (5 mg/kg). Rats were allowed to recover for at least 8 days.

Correct placement of the cannulae and of electrodes for LFP recordings was confirmed by histology after completion of the experiments. For this, the rats were perfused intracardially with 0.9% saline followed by 4% paraformaldehyde (PFA). After decapitation, the brains were removed and immersed in the 4% PFA for at least two days. Coronal sections of 50–70 µm were cut on a vibratome, stained with toluidine blue and examined under a light microscope (Extended Data Fig. 5).

Analysis of memory performance. Exploration was defined by the rat being within 2 cm of an object, directing its nose towards the object and engaging in active exploration behaviours such as sniffing. For each task, the time a rat spent exploring each object during the retrieval test was converted into a discrimination ratio according to the general formula: (time spent at novel − time spent at familiar)/(time spent at novel + time spent at familiar), where 'novel' on the NOR task refers to the novel object and on the OPR task refers to the displaced object. A value of zero indicates no exploration preference, whereas a positive value indicates preferential exploration of the novel configuration, thus indicating memory of the familiar configuration. Additionally, the total time of object exploration (across both objects), distance travelled and mean speed on each task were determined. Statistical comparisons concentrated on cumulative discrimination ratios for the first 1 min and 3 min of the retrieval phase.

Analysis of sleep, EEG, and hippocampal LFP recordings. Sleep during the retention interval was assessed using video recordings and tracking software (ANY-Maze, Stoelting Europe) using standard visual procedures⁴¹. In brief, sleep was scored whenever the rat showed a typical sleep posture and stayed immobile for at least 10 s. If brief movements interrupted sleep epochs by <5 s, continuous sleep was scored. The agreement of the procedure with EEG-based scoring of sleep in the present (see below) and previous studies was >92%^{22,41}. Scores indicated an average of 46.97 ± 2.86 min spent asleep during the 2-h post-encoding retention interval, with the first bout of sleep occurring 41.24 ± 2.99 min after the encoding phase. There were no significant differences in sleep parameters between NOR and OPR task conditions or retention intervals tested (Extended Data Table 1).

In the experiments testing the effects of reversible inactivation of the hippocampus, sleep was additionally analysed using EEG and EMG recordings. For the recordings, electrodes were connected through a preamplifier headstage (Model HS-18MM, Neuralynx) to a Digital Lynx SX acquisition system (Neuralynx),

amplified, filtered (EEG: 0.01–300.0 Hz; EMG: 30.0–300.0 Hz), and sampled at a rate of 1,000 Hz. Sleep stages (SWS, preREM and REM sleep) and wakefulness were scored offline by visual inspection using 10-s epochs according to standard criteria⁴². In brief, the wake stage was characterized by predominant low-amplitude fast activity associated with increased EMG tonus. SWS was characterized by predominant high-amplitude delta activity (<4.0 Hz) and reduced EMG activity, and REM sleep by predominant theta activity (4.0–8.0 Hz), phasic muscle twitches and minimal EMG activity. PreREM sleep was identified by a decrease in delta activity, a progressive increase in theta activity and the presence of sleep spindles (10.0–16.0 Hz). Sleep stage classification was performed by an experienced experimenter.

EEG signals in these experiments were also used to identify slow oscillations and spindles during SWS. Identification of slow oscillations followed procedures as described¹⁷. In brief, the EEG signal during all SWS epochs for an animal was filtered between 0.3 and 4.5 Hz. A slow oscillation event was then identified if the following criteria were fulfilled: (i) two consecutive negative-to-positive zero crossings of the signal occurred at an interval between 0.4 and 2.0 s; (ii) of these events in an individual rat, the 35% with the highest negative peak amplitude between both zero crossings were selected; and (iii) of these events the 45% with the highest negative-to-positive peak-to-peak amplitude were selected. These criteria resulted in the detection of slow oscillations with negative peak amplitudes exceeding $-80 \mu\text{V}$ and peak-to-peak amplitudes exceeding $120 \mu\text{V}$. For spindle detection, the EEG signal was filtered between 10.0 and 16.0 Hz. The Hilbert transform was calculated for the filtered signal and smoothed with a moving average (window size 200 ms). A spindle was identified when the absolute value of the transformed signal exceeded 1.5 s.d. of the mean signal during the animal's SWS epochs, for at least 0.4 s and not more than 2.0 s.

The same procedures were applied to identify slow oscillations and spindles in the hippocampal LFP recordings. To identify ripples in these LFP recordings, the signal was filtered between 150.0 and 250.0 Hz. As for spindle detection, the Hilbert transform was calculated and the signal was smoothed using a moving average (window size 200 ms). A ripple event was identified when the Hilbert transform value exceeded a threshold of 2.5 s.d. from the mean signal during an animal's SWS epochs, for at least 25 ms (including at least 3 cycles) and for not more than 500 ms. **Statistical analyses.** Statistical analyses were performed using SPSS 21.0 for Windows. To evaluate the discrimination ratios determined for each task, we used ANOVAs that included group factors for the task (NOR/OPR) or the

retention interval (1/3 weeks), and repeated-measures factors representing the sleep/wake conditions and discrimination ratios after 1 and 3 min of the retrieval phase. (ANOVAs separately run on 1-min and 3-min values yielded almost identical results and are not reported here.) Muscimol/vehicle comparisons were introduced as group or repeated-measures factors, depending on the experiment. ANOVAs indicating significance for main or interaction effects of interest were followed by post hoc *t*-tests (two-sided). Discrimination ratios were also compared with chance level performance (zero) using one-sample *t*-tests. To analyse the relationship between post-encoding retention sleep and memory performance, Pearson product-moment correlation coefficients were calculated. $P < 0.05$ was considered significant.

Code availability. The codes used in this study are available from the corresponding authors on reasonable request. MATLAB scripts used for analyses of EEG and LFP signals are available at https://github.com/MedPsych/LongTermMemory_Sleep.

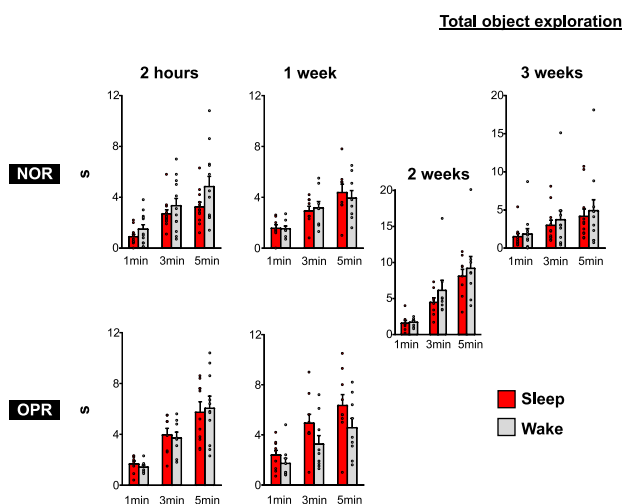
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

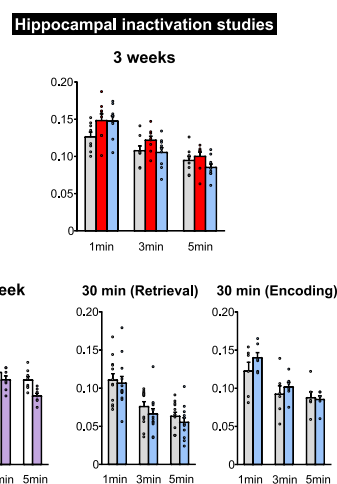
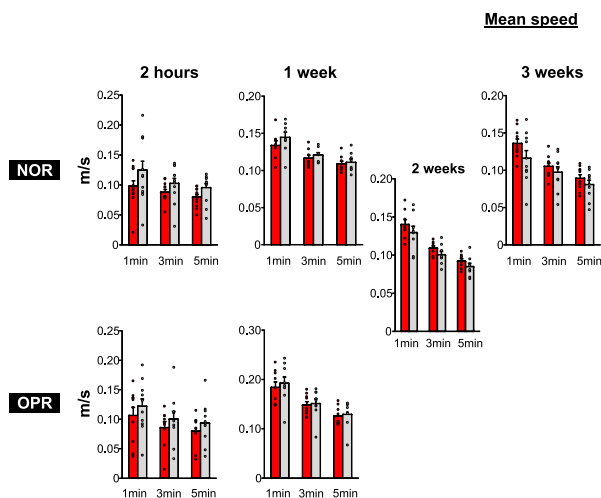
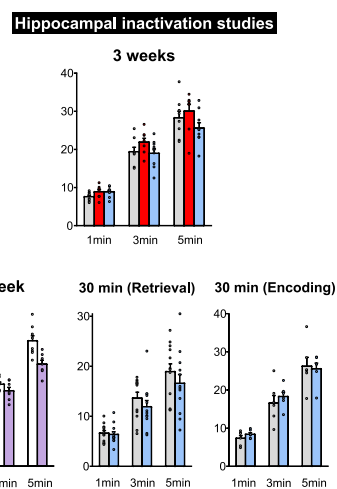
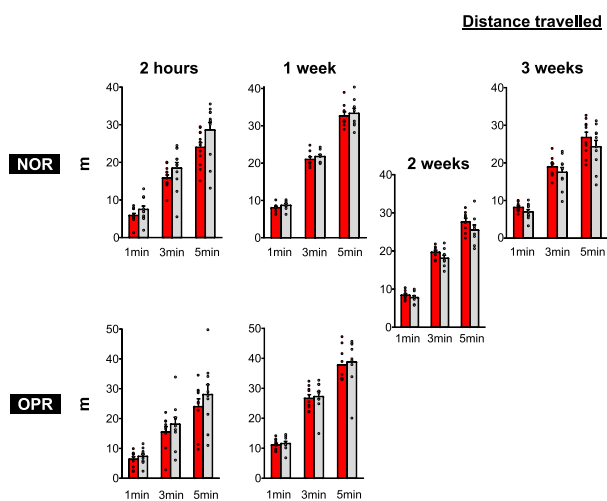
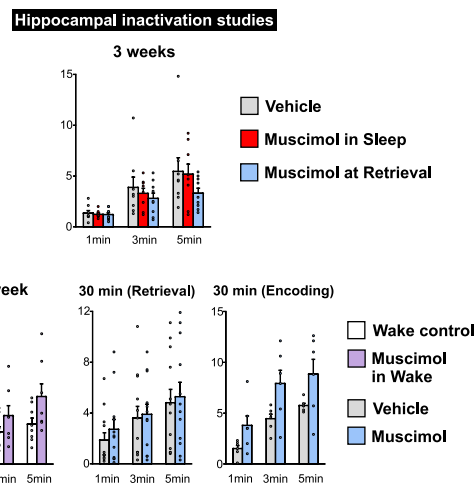
The data that support the findings of this study are available from the corresponding authors on reasonable request. Source Data for graphs shown in Figs. 1–3 and Extended Data Figs. 1–4 are available in the online version of the paper.

43. Dix, S. L. & Aggleton, J. P. Extending the spontaneous preference test of recognition: evidence of object-location and object-context recognition. *Behav. Brain Res.* **99**, 191–200 (1999).
44. Chambon, C., Wegener, N., Gravius, A. & Danysz, W. A new automated method to assess the rat recognition memory: validation of the method. *Behav. Brain Res.* **222**, 151–157 (2011).
45. Ozawa, T., Yamada, K. & Ichitani, Y. Long-term object location memory in rats: effects of sample phase and delay length in spontaneous place recognition test. *Neurosci. Lett.* **497**, 37–41 (2011).
46. Goshen, I. et al. Dynamics of retrieval strategies for remote memories. *Cell* **147**, 678–689 (2011).
47. Allen, T. A. et al. Imaging the spread of reversible brain inactivations using fluorescent muscimol. *J. Neurosci. Methods* **171**, 30–38 (2008).
48. Bonnevie, T. et al. Grid cells require excitatory drive from the hippocampus. *Nat. Neurosci.* **16**, 309–317 (2013).

a



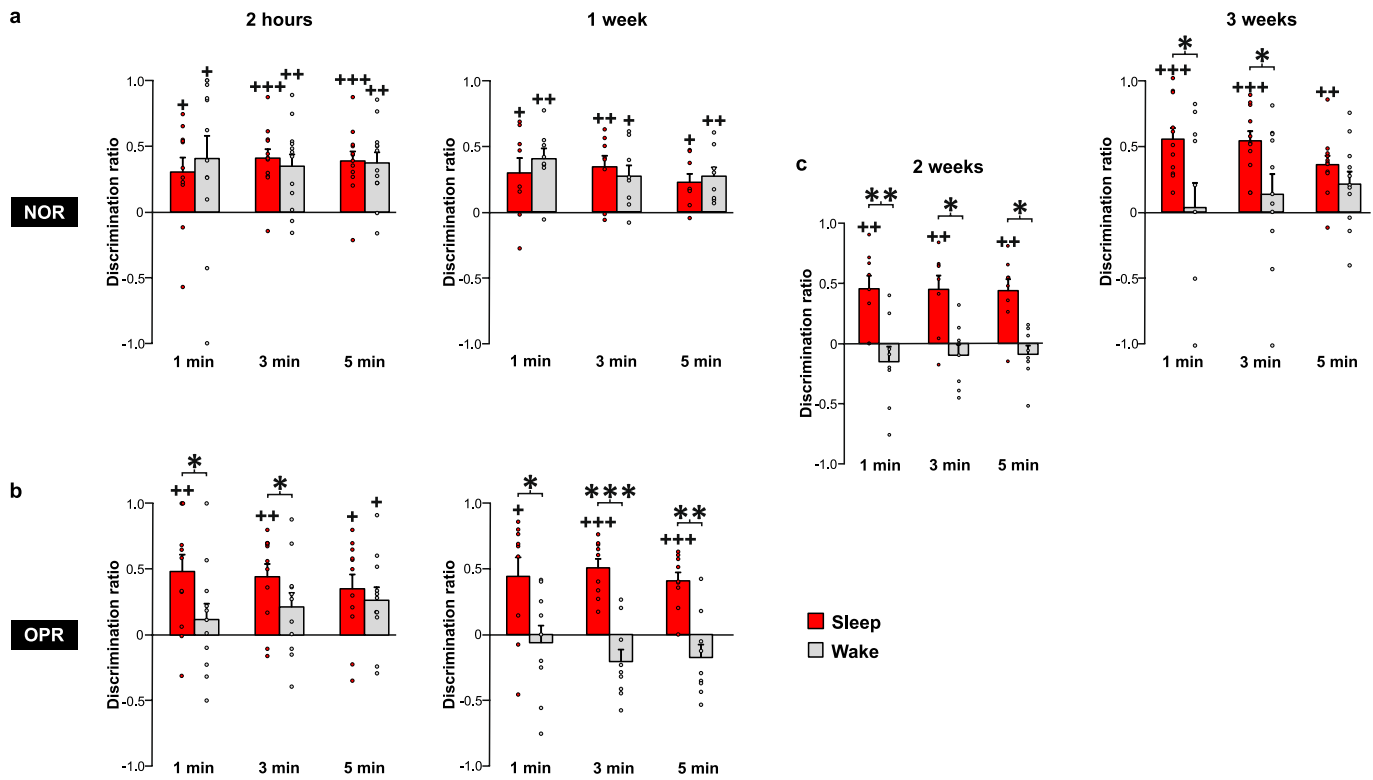
b



Extended Data Fig. 1 | See next page for caption.

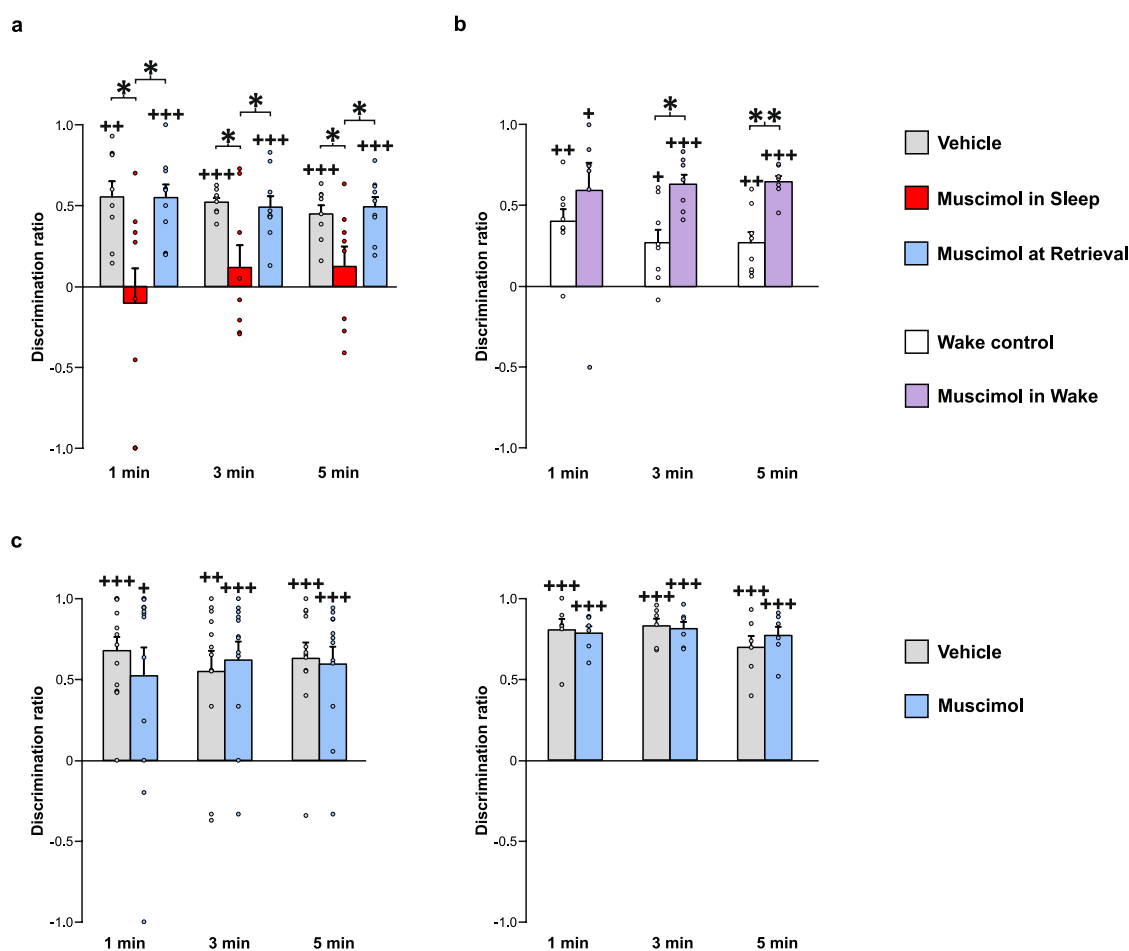
Extended Data Fig. 1 | Control measures for NOR and OPR task performance. Total object exploration (s), total distance travelled (m) and average speed (m s^{-1}) at retrieval testing. Mean values (\pm s.e.m., dot plots overlaid) for the first 1 and 3 min and for the entire 5 min of the retrieval phase are shown. **a**, Results from main experiments of NOR and OPR memory as illustrated in Fig. 1. Retrieval was tested either immediately after the 2-h retention interval (recent) or 1 week or (for the NOR task only) 3 weeks later (remote). In a supplementary experiment, NOR was tested 2 weeks after encoding (offset downwards). Red, sleep; grey, wake; $n = 12$, 8, 8 and 11 rats for NOR testing after 2 h and 1, 2 and 3 weeks, and $n = 11$ and 9 rats for OPR testing after 2 h and 1 week, respectively. **b**, Results from experiments after bilateral intrahippocampal infusion of muscimol as in Fig. 2. Top, muscimol (versus vehicle, grey bars, $n = 8$ rats) was infused either during the 2-h post-encoding interval (upon

first occurrence of SWS, red bars, $n = 8$ rats) or 15 min before retrieval (blue bars, $n = 9$ rats) with the retrieval phase taking place 3 weeks after encoding. Bottom, control studies. Left, muscimol (purple, $n = 7$ rats) was infused shortly after encoding while the rats remained awake during the 2-h post-encoding interval, compared with untreated wake control rats ($n = 8$ rats, empty bars). Retrieval was tested 1 week after encoding (corresponding to Fig. 2b). Right, muscimol (blue bars, versus vehicle, grey bars) was infused either 15 min before retrieval testing ($n = 12$ rats) or 15 min before encoding ($n = 6$ rats) with the retrieval phase taking place 30 min after encoding (corresponding to Fig. 2c). There were no significant differences between sleep and wake or between muscimol and vehicle conditions ($P > 0.194$, for all comparisons based on ANOVA and two-sided post hoc t -tests, see Methods and Figs. 1, 2 for further details).



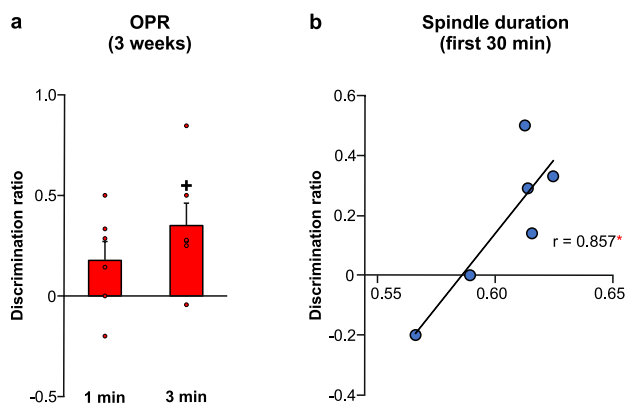
Extended Data Fig. 2 | Performance in recent and remote tests for NOR and OPR tasks. Memory is indicated by mean \pm s.e.m. discrimination ratios during the first 1 min, first 3 min, and entire 5 min of the retrieval phase on the NOR and OPR tasks (dot plots overlaid). **a**, NOR was tested with 2-h (recent) and with 1-week and 3-week (remote) retrieval tests. **b**, OPR was tested with 2-h (recent) and 1-week retrieval tests. Whereas OPR memory benefited from sleep (red bars; compared to wake, grey) at both recent and remote (1 week) retrieval tests, NOR benefited from sleep only at the 3-week retrieval test, when NOR memory had decayed in the wake condition. **c**, A supplementary experiment with NOR retrieval tested 2 weeks after post-encoding sleep and wake intervals showed that NOR memory in the wake condition had already faded at this 2-week point, whereas it was preserved in the sleep condition ($F_{1,7} = 14.997$, $P = 0.006$, for sleep/wake main effect; $F_{1,14} = 18.151$, $P = 0.01$ and

$F_{1,14} = 0.82$, $P = 0.382$, for 1 versus 2-week comparisons in the wake and sleep conditions, respectively, $F_{1,14} = 12.073$, $P = 0.005$, for 1/2 weeks \times sleep/wake interaction; $P > 0.222$ for all comparisons between 2- and 3-week retrieval). In all experiments, recognition memory was assessed by the discrimination ratios during the first 1 and first 3 min of the retrieval period, which typically cover exploration of novelty most sensitively on both the NOR and OPR tasks^{6,43–45}. With extended exploration periods, the novelty response often decreases and is thought to become more noisy. Hence, here, the 5-min values were not used for the assessment of recognition memory. $n = 12, 8, 11$ and 8 rats for NOR testing at 2 h, 1 week, 3 weeks and 2 weeks; $n = 11$ and 9 rats for OPR testing at 2 h and 1 week, respectively. +++ $P < 0.001$, ++ $P < 0.01$, + $P < 0.05$ for one-sample t -test against chance level; *** $P < 0.001$, * $P < 0.05$ for pairwise t -tests (two-sided) between sleep and wake conditions.

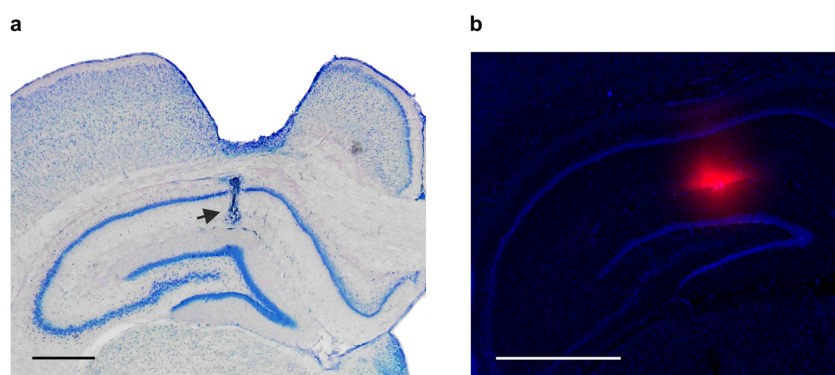


Extended Data Fig. 3 | Performance on NOR task for hippocampal inactivation studies. Memory is indicated by mean \pm s.e.m. discrimination ratios during the first 1 min, first 3 min, and entire 5 min of the retrieval phase on the NOR task in experiments involving reversible inactivation of the hippocampus (dot plots overlaid). **a**, Muscimol (red bars, $n = 8$ rats, versus vehicle, grey bars, $n = 8$ rats) was infused into the hippocampus in the post-encoding interval upon the first occurrence of continuous SWS, or 15 min before retrieval testing (blue bars, $n = 9$ rats). Retrieval was tested 3 weeks after encoding. **b**, Control study in which muscimol (purple bars, $n = 7$ rats) was infused shortly after encoding while the rats remained awake during the 2-h post-encoding interval, compared with untreated wake control rats ($n = 8$ rats, empty bars). Retrieval was tested 1 week after encoding. Infusion of muscimol during

post-encoding wakefulness tended to enhance NOR performance, which suggests that during wakefulness hippocampal activity normally interferes with NOR memory consolidation⁸. It might also reflect compensatory plasticity occurring in extrahippocampal regions upon hippocampal suppression⁴⁶. **c**, Control studies in which muscimol (blue bars, versus vehicle, grey bars) was infused 15 min before retrieval testing of recent NOR memory (left, $n = 12$ rats for each substance condition) or 15 min before the encoding phase (right, $n = 6$ rats for each substance condition). Retrieval was tested 30 min after encoding, with the rats staying awake during this interval. $+++P < 0.001$, $++P < 0.01$, $+P < 0.05$ for one-sample t -test against chance level; $**P < 0.01$, $*P < 0.05$ for pairwise t -tests (two-sided) between conditions. See Fig. 2 for further details.



Extended Data Fig. 4 | Remote 3-week OPR testing. OPR memory was tested in $n = 6$ rats, 3 weeks after a 2-h post-encoding sleep interval. These supplementary experiments followed the same procedures as described for the 1-week sleep condition on the OPR task, but included sleep EEG recordings. **a**, OPR memory is indicated by the mean \pm s.e.m. discrimination ratio during the first 1 min and 3 min of exploration. $^+P = 0.034$, for one-sample t -test against chance level. Rats displayed significant OPR memory after 3 min (as well as for the whole 5-min exploration period). **b**, OPR performance (discrimination ratio at 1 min) at the 3-week retrieval test was correlated with sleep spindle duration during the first 30 min of post-encoding sleep ($^*P = 0.029$, Pearson's product-moment correlation). A similar correlation with NOR performance at the 3-week retrieval (Fig. 3a) points towards a similar mechanism underlying the formation of long-term NOR and OPR memory during sleep.



Extended Data Fig. 5 | Verification of cannula location and muscimol spreading. **a**, Coronal brain section showing location of cannula in the dorsal hippocampus (black arrow) together with position of guide cannula in overlying cortex. **b**, Coronal brain section showing spread of muscimol (red) after infusion into the hippocampus. Experiments were repeated in $n = 3$ rats with similar results. The infusion protocol was the same as in the behavioural experiments. In brief, after implantation of the

guide cannula in the dorsal hippocampus, animals were infused using the injection cannulae with $0.5 \mu\text{l}$ fluorophore-conjugated muscimol^{47,48}. After infusion, animals were intracardially perfused and brains were post-fixed with PFA 4% for 24 h. Brains were cut on a vibratome to obtain $70\text{-}\mu\text{m}$ -thick sections and stained with DAPI ($1:5,000 \mu\text{l}$ in PBS) for 15 min. Fluorescent images were acquired by epifluorescence microscopy (Axio imager Zeiss, Germany). Scale bars, 1 mm.

Extended Data Table 1 | Sleep parameters

a

Sleep parameter	NOR			OPR	
	2 hours	1 week	3 weeks	2 hours	1 week
Duration (min)	56.87 ± 6.36	53.12 ± 8.92	41.85 ± 7.10	46.94 ± 5.63	40.05 ± 5.01
Latency (min)	31.46 ± 5.77	29.78 ± 8.06	43.24 ± 6.35	44.57 ± 6.73	43.24 ± 4.06

b

Sleep parameter	Latency (min)		Duration (min)	
	SWS	SWS	PreREM	REM
Vehicle	20.80 ± 5.71	47.13 ± 5.44	5.80 ± 0.71	6.04 ± 1.07
Muscimol	17.57 ± 6.16	53.63 ± 11.15	1.97 ± 0.51**	0.83 ± 0.83**

c

SWS parameter	SO density (number/min)	SO amplitude (mV)	Spindle density (number/min)	Spindle power (mV ² /s)	Spindle mean duration (s)
Vehicle	32.61 ± 4.06	0.189 ± 0.021	2.81 ± 0.15	0.026 ± 0.002	0.591 ± 0.011
Muscimol	31.62 ± 3.47	0.187 ± 0.015	2.86 ± 0.10	0.025 ± 0.002	0.598 ± 0.022

a, Sleep duration and latency during the 2-h post-encoding interval for the sleep groups of the main experiments (Fig. 1). In these experiments retrieval was tested either immediately after the 2-h retention interval (test of recent memory) or 1 week or (for the NOR task only) 3 weeks later (tests of remote memory). There were no significant differences between NOR and OPR task conditions or retention intervals. $n = 12, 8$, and 11 rats for NOR testing after 2 h, 1 week and 3 weeks, and $n = 11$ and 9 rats for OPR testing after 2 h and 1 week, respectively. **b**, Post-encoding sleep in the experiments after bilateral intrahippocampal infusion of muscimol (Fig. 2a). Sleep latency, time in SWS, preREM sleep, and REM sleep are indicated ($n = 8$ rats for each condition). **c**, For the same experiments, density and amplitude of slow oscillations (SO) and density, power, and mean duration of spindles identified during SWS are indicated for the vehicle and muscimol conditions. Substances were infused during the 2-h post-encoding interval (upon the first occurrence of SWS). **PreREM $P = 0.002$, REM $P = 0.003$, for pairwise t-tests (two-sided) with vehicle condition. Data shown as mean ± s.e.m.

Extended Data Table 2 | Correlations between NOR after 3 weeks and sleep parameters

	SWS	Spindles				Slow Oscillations			
	Duration	Number	Duration	Density	Power	Number	Duration	Density	Power
<i>r</i>	0.536	0.719	0.705	0.654	-0.192	0.259	-0.215	-0.132	-0.319
<i>P</i>	0.137	0.029*	0.034*	0.056	0.620	0.501	0.578	0.735	0.402

REM Sleep		
	Duration	Theta power
<i>r</i>	0.251	0.081
<i>P</i>	0.514	0.836

Summary of correlations between NOR performance at the 3-week retrieval (1 min discrimination ratio) and sleep parameters during the 2-h post-encoding interval ($n = 9$ rats). Pearson's correlation coefficients and P values are indicated. * $P < 0.05$ level (uncorrected).

Mechanosignalling via integrins directs fate decisions of pancreatic progenitors

Anant Mamidi^{1,3}, Christy Prawiro^{1,3}, Philip A. Seymour¹, Kristian Honnens de Lichtenberg¹, Abigail Jackson¹, Palle Serup¹ & Henrik Semb^{1,2*}

The pancreas originates from two epithelial evaginations of the foregut, which consist of multipotent epithelial progenitors that organize into a complex tubular epithelial network. The trunk domain of each epithelial branch consists of bipotent pancreatic progenitors (bi-PPs) that give rise to both duct and endocrine lineages, whereas the tips give rise to acinar cells¹. Here we identify the extrinsic and intrinsic signalling mechanisms that coordinate the fate-determining transcriptional events underlying these lineage decisions^{1,2}. Single-cell analysis of pancreatic bipotent pancreatic progenitors derived from human embryonic stem cells reveal that cell confinement is a prerequisite for endocrine specification, whereas spreading drives the progenitors towards a ductal fate. Mechanistic studies identify the interaction of extracellular matrix (ECM) with integrin $\alpha 5$ as the extracellular cue that cell-autonomously, via the F-actin–YAP1–Notch mechanosignalling axis, controls the fate of bipotent pancreatic progenitors. Whereas ECM–integrin $\alpha 5$ signalling promotes differentiation towards the duct lineage, endocrinogenesis is stimulated when this signalling cascade is disrupted. This cascade can be disrupted pharmacologically or genetically to convert bipotent pancreatic progenitors derived from human embryonic stem cells to hormone-producing islet cells. Our findings identify the cell-extrinsic and intrinsic mechanotransduction pathway that acts as gatekeeper in the fate decisions of bipotent pancreatic progenitors in the developing pancreas.

To investigate whether cell-autonomous or non-autonomous mechanisms regulate the expression of the fate-determining transcription factor PDX1, we differentiated human embryonic stem cells (hESCs) expressing PDX1–GFP into pancreatic progenitors including bipotent pancreatic progenitors (bi-PPs)^{3,4} (Extended Data Fig. 1a–e). Sorted GFP^{high} cells were able to attach as single cells and self-aggregate (Extended Data Fig. 2a). Whereas cells in the central region of the cell clusters maintain high PDX1 expression (Extended Data Fig. 2b, c), spread cells at the periphery of the aggregates and spread single cells downregulate PDX1 expression (Extended Data Fig. 2b–d). Notably, single confined cells maintained high PDX1 expression, suggesting that cell shape may govern expression of PDX1 in a cell-autonomous manner. Numerous studies have demonstrated that changes in cell geometry and spreading have profound effects on cell fate decisions^{5–7}. To study the effect of cell geometry and spreading on fate decisions of single PDX1^{high} pancreatic progenitors, we made use of micropatterned glass slides (Fig. 1a). Consistent with the self-aggregation data, physical restraint of spreading (confinement) maintained PDX1^{high} expression, whereas cell spreading resulted in reduced PDX1 expression (Fig. 1a, Extended Data Fig. 2e, f). Cell geometry (disc or square) had no effect on PDX1 expression (Extended Data Fig. 2g). Analysis of other pancreatic transcription factors showed a similar response for NKX6.1 expression, whereas the expression of SOX9, FOXA2 and HNF6 were unaffected by the size of the adhesion area (Fig. 1a, Extended Data Fig. 2g).

Endocrinogenesis in the developing human and mouse pancreas is accompanied by a gradual reduction of the mechanoresponsive

transcription factor YAP1^{8–10}. Cell spreading maintains nuclear YAP1 activity, whereas cell confinement (below 500 μm^2) leads to loss of YAP1 expression (Fig. 1a, b Extended Data Fig. 2i, j) and induction of the endocrine progenitor marker NEUROG3 (also known as NGN3) (Fig. 1a). YAP1 levels are significantly reduced at both mRNA and protein level in hESC-derived NGN3⁺ cells (Extended Data Fig. 3e). On the basis of these data, we proposed that YAP1-mediated mechanical influences transcriptionally regulate the fate choice of bi-PPs (YAP1⁺PDX1^{high}) into duct (YAP1⁺PDX1^{low}) or endocrine (YAP1⁺NGN3⁺) lineages (Fig. 1c).

YAP1 and its DNA-binding co-factors TEAD1–TEAD4, and their target HES1, are highly expressed during early stages of mouse pancreas development⁹ (embryonic day (E)9–12.5), whereas their expression decreases when differentiation of the endocrine lineages is peaking^{9,10} (Extended Data Fig. 3a–d). These data support and extend the previous findings that endocrinogenesis is associated with reduced expression of YAP1 and TEAD1–TEAD4^{9,10}. To validate the findings from single cells in vivo, we deleted YAP1 in the pancreatic epithelium using *Pdx1-cre; Yap1^{fl/fl}* mice^{11,12}. This leads to hypoplasia and hypoglycaemia (Fig. 2a, Extended Data Fig. 3f), which can be explained by increased endocrinogenesis (Fig. 2a, b, Extended Data Fig. 3h). Using a tamoxifen-inducible Cre driver (*Sox9-creER^{T2}; Yap1^{fl/fl}*)¹³ we corroborated these results (Extended Data Fig. 3g, i) and also showed that the increase in endocrinogenesis is associated with a depletion of bi-PPs (Extended Data Fig. 3i). Notably, no compensatory increased expression of *Taz* was observed upon *Yap1* ablation (Extended Data Fig. 3i). A corresponding increase in endocrinogenesis was observed when mouse E11.5 explants were treated with verteporfin, an inhibitor of YAP1–TEAD complex formation¹⁴ (Extended Data Figs. 3j, 4d). These results indicate that loss of YAP1 in vivo directs bi-PPs to the endocrine lineage. Inhibiting YAP1 function (using short interfering RNA (siRNA) or verteporfin) in pancreatic progenitors derived from hESCs resulted in increased NGN3 and insulin protein levels together with increased levels of *NEUROG3*, *NEUROD1*, *ISL1*, *GCG*, *INS* and *MAFB* mRNAs (Fig. 2c, d and Extended Data Fig. 4a–c). This showed that YAP1 has a conserved role as a repressor of endocrinogenesis in humans. In sum, these results indicate that YAP1 maintains bi-PPs by inhibiting their differentiation into the endocrine lineage, suggesting that the extracellular cue that inactivates YAP1 may also be the trigger of endocrinogenesis.

The results of single-cell analysis suggest that maintenance of YAP1 expression in spread cells correlates with a duct phenotype (Extended Data Fig. 2d, e, h), which is consistent with the expanded ductal compartment that is observed upon constitutive (*Rosa26*-promoter driven) expression of YAP1 in pancreatic progenitors¹⁵ (at E17.5). We confirmed and extended these findings by expressing constitutively active human YAP1(S127A) in mouse embryos at an earlier time-point (E12.5, *tet-YAP1^{S127A}; Pdx1^{TA/+}*, hereafter referred to as YAP1tg mice^{16,17}). Control experiments showed that expression of YAP1(S127A) upregulated expression of known downstream YAP1

¹Novo Nordisk Foundation Center for Stem Cell Biology (DanStem), University of Copenhagen, Copenhagen, Denmark. ²Institute of Translational Stem Cell Research, Helmholtz Zentrum München, Neuherberg, Germany. ³These authors contributed equally: Anant Mamidi, Christy Prawiro. *e-mail: semb@sund.ku.dk

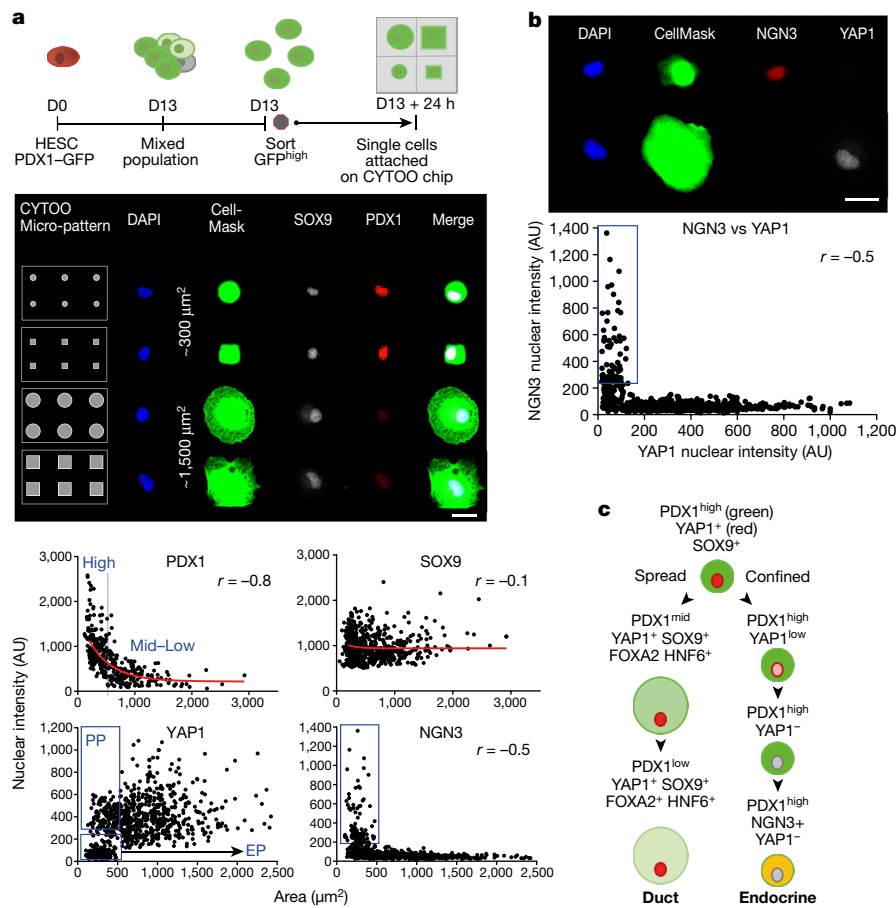


Fig. 1 | Cell spreading determines cell specification and gene expression in pancreatic progenitors. **a**, Single pancreatic progenitor cells derived from hESCs sorted on micropatterned slides 24 h after adhesion. Top, schematic of single pancreatic progenitor cells sorted on a micropatterned slide. Middle, each micropattern contains a single pancreatic progenitor cell. Representation of circular and square confined pancreatic progenitor cells (top two micropatterns, respectively) and spread pancreatic progenitor cells (bottom two micropatterns, respectively). DAPI, blue; CellMask, green; SOX9, grey; PDX1, red. Scale bar, 20 μm . Bottom, nuclear intensity of pancreatic markers plotted against cell area. Each data point corresponds to an individual cell. PDX1, $n = 369$; SOX9, $n = 593$; YAP1, $n = 906$; NGN3, $n = 906$; 3 independent experiments. Spearman's correlation coefficient (r) is calculated between cell area and nuclear intensity for: PDX1, $r = -0.8$ ($P \leq 0.0001$); SOX9, $r = -0.1$; YAP1, $r = 0.5$ ($P \leq 0.0001$); NGN3, $r = -0.5$ ($P \leq 0.0001$). AU, arbitrary units. **b**, Correlation of expression (r) between NGN3 and YAP1. DAPI, blue; NGN3, red; YAP1, grey; CellMask, green. Scale bar, 20 μm . NGN3 and YAP1 are negatively correlated; $r = -0.5$ ($P \leq 0.0001$). Each data point corresponds to an individual cell; $n = 906$ (NGN3, YAP1); 3 independent experiments. **c**, Model of progression of single pancreatic progenitor cell fate specification during spreading or confinement over 24 h.

target genes, such as *Cdc20*, *Ctcf* (also known as *Ccn2*), *Birc5* (also known as baculoviral IAP repeat containing 5) and *Snai2* (Extended Data Fig. 4f, g). Whereas the size of the early YAP1tg pancreas (E12.5) does not appear changed compared to controls, perturbed branching and tip and trunk patterning defects were apparent, as indicated by expression of SOX9, HNF1 β , MUC1 and E-cadherin, and staining with the duct-specific lectin DBA (*Dolichos biflorus* agglutinin), throughout the unbranched epithelium at E12.5 and E15.5 (Fig. 3c, Extended Data Fig. 5b, Supplementary Video 1). Furthermore, the reduced expression of pancreatic progenitor (*Pdx1* and *Nkx6-1*), acinar (*Cpa1* and *Amy2a5*) and endocrine (*Ngng3* (also known as *Neurog3*) and *Ins1*) markers (Extended Data Figs. 4h, 5a, c) and upregulation of ductal genes such as *Sox9* and *Hnf1b* (Extended Data Fig. 5b, c) show that progenitor maintenance and differentiation are also aberrant in the YAP1tg embryos. Subsequently, the YAP1tg embryos exhibit severe pancreas agenesis (E18.5), and newborn pups die as a result of severe hyperglycaemia within a week of birth (Extended Data Fig. 4e). Together with the single-cell analysis (Extended Data Fig. 2h), these results indicate that sustained or increased expression of YAP1 within bi-PPs triggers specification to the duct lineage (Fig. 3c, Extended Data Fig. 5b, Supplementary Video 1).

Several signal transduction pathways have been identified as transcriptional targets of YAP1, including Notch, TGF β , BMP, MAPK, canonical Wnt and mTOR pathways^{9,18}. HES1 and Notch1 are upregulated in YAP1tg pancreata at E15.5 (Extended Data Fig. 5c), and YAP1tg phenocopies overexpression of the Notch1 intracellular domain (NICD) in the developing pancreas^{19,20}, strongly suggesting that in the developing pancreas YAP1 acts upstream of Notch1. This is unlike its activity in epidermal stem cells, in which YAP1 is regulated by active Notch signalling²¹. Blocking Notch signalling (using a γ -secretase inhibitor) in YAP1tg explants reduces expression of *Hes1* mRNA and partially restores expression of both acinar (*Ptf1a*, *Cpa1* and *Amy2a5*) and endocrine (*Ngng3*, *Ins1* and *Gcg*)

markers (Fig. 3d, Extended Data Fig. 5d). Consistent with previous results from chromatin immunoprecipitation followed by sequencing (ChIP-seq) using TEAD1⁹, we also show that YAP1 and TEAD4 bind specifically to *Hes1* and *NOTCH1* loci in hESC-derived pancreatic progenitors (Fig. 3a). Together, these data suggest that YAP1-mediated transcriptional repression of *Ngng3* is mediated by its transcriptional activation of *Hes1*. Furthermore, ENCODE ChIP-seq data²², together with analysis by chromatin immunoprecipitation followed by quantitative PCR, demonstrate that YAP1, TEAD4 and HES1 specifically bind to the *NGN3* promoter in pancreatic progenitors (Fig. 3b, Extended Data Fig. 6a). To assess the potential co-repressor role of YAP1-TEAD4^{23,24} on the *NGN3* promoter (800 bp upstream of the promoter), we carried out luciferase assays in hESC-derived pancreatic progenitors. Reduced expression of YAP1 activates the human *NGN3* promoter, whereas constitutive expression of either YAP1, rat HES1 or rat NICD partially restores YAP1-mediated transcriptional repression of *NGN3* (Extended Data Fig. 6b). Furthermore, YAP1 expression partially prevents the transcriptional activation of *NGN3* upon siRNA-mediated knockdown of *HES1* mRNA (Extended Data Fig. 6b). As expected, mutating the HES1 binding sites on the *NGN3* promoter leads to increased *NGN3* transcription, whereas mutational inactivation of the TEAD binding site failed to affect *NGN3* transcription (Extended Data Fig. 6b). Of note, co-immunoprecipitation experiments demonstrate that YAP1, HES1 and TEAD4 physically interact with each other in pancreatic progenitors (Extended Data Fig. 6c). To exclude the possibility that YAP1 acts downstream of Notch, we show that nuclear localization of YAP1 is maintained in SOX9⁺ pancreatic progenitors of *Foxa2*^{T2Aicre}; *Rosa26*^{dnMaml1-eGFP} mutant²⁵ mice, in which Notch signal transduction is blocked, at E10.5 and E15.5 (Extended Data Fig. 5e, f). Collectively, these data implicate a dual role of YAP1 in endocrinogenesis, both as an activator of *HES1* transcription and as a co-repressor of *NGN3* transcription through chromatin looping of distant YAP1-TEAD binding sites (Extended Data Fig. 6c).

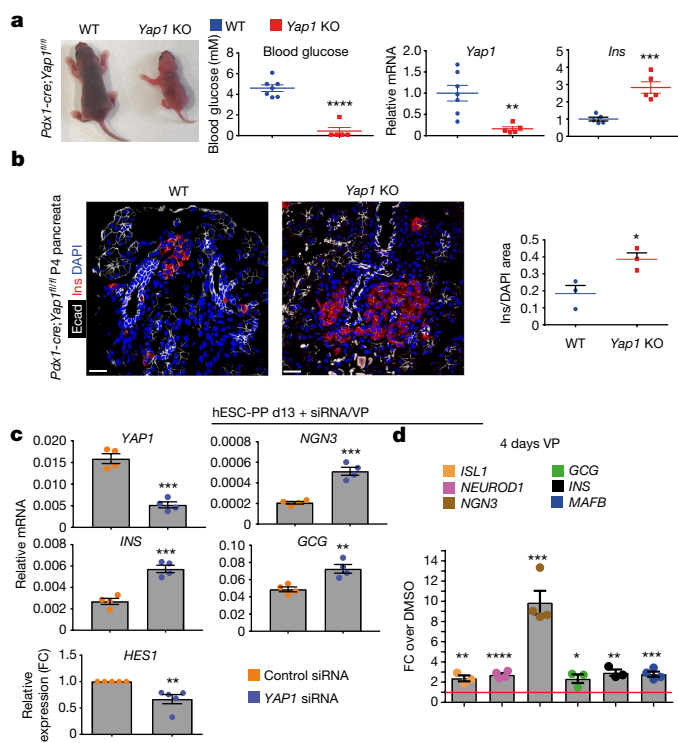


Fig. 2 | YAP1 deletion in the pancreatic progenitor initiates endocrinogenesis both in vivo and in vitro. **a**, *Pdx1*-Cre-mediated *Yap1* knockout (*Yap1* KO) results in hypoglycaemia at post-natal day 4. Random fed blood glucose measurements. Wild type, $n = 7$; *Yap1* KO, $n = 5$; **** $P < 0.0001$. Relative gene expression analysis by quantitative reverse transcription with PCR (qRT-PCR) in P4 pancreata for *Yap1* (wild type, $n = 7$; *Yap1* KO, $n = 5$; ** $P = 0.0031$) and insulin (*Ins1*) (wild type, $n = 5$; *Yap1* KO, $n = 5$; *** $P = 0.0022$). Two-tailed unpaired *t*-test; mean \pm s.d. **b**, Immunostaining of E-cadherin (Ecad, grey) and insulin (Ins, red) in P4 control and *Yap1* KO pancreata. DAPI, blue. Scale bar, 26 μ m. Right, quantifications of insulin⁺ area/DAPI⁺ area (wild type, $n = 3$; *Yap1* KO, $n = 3$; * $P = 0.0295$). Two-tailed unpaired *t*-test. **c**, Pancreatic progenitor cells derived from hESCs were transfected with *YAP1* siRNA on day 13 (d13) and fixed after 72 h. Gene expression was measured by qRT-PCR. Data are mean \pm s.e.m. *YAP1*, *NGN3*, *INS1*, $n = 4$; *HES1*, $n = 5$. ** $P \leq 0.01$, *** $P \leq 0.001$; two-tailed unpaired *t*-test. FC, fold change; VP, verteporfin. **d**, Day 13 cells expressing PDX1-GFP were treated with verteporfin for 2 days followed by 2 days without the drug and analysed for gene expression by qRT-PCR; * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$, **** $P \leq 0.0001$; two-tailed unpaired *t*-test; data are mean \pm s.d.

Cell spreading positively regulates YAP1 activity through the formation of actin bundles, whereas cell confinement negatively regulates YAP1 activity by promoting dissociation of actin bundles²⁶. Similarly, highly confined single NGN3⁺ cells (less than 500 μ m²) exhibited reduced stress fibre formation (Fig. 4a). This observation was substantiated by the reduced levels of F-actin observed in NGN3⁺ endocrine precursor cells compared to pancreatic progenitors in vivo and in vitro (Fig. 4b, Extended Data Fig. 7a). Treatment with latrunculin B (latB), a small-molecule inhibitor that blocks YAP1 activity²⁷ by disassembling F-actin bundles through sequestering G-actin²⁸, reduced cell spreading and nuclear localization of YAP1 with a concomitant upregulation of NGN3 expression in single pancreatic progenitors (Fig. 4c, Extended Data Fig. 8a). Moreover, treating unsorted hESC-derived pancreatic progenitors and mouse pancreatic explants with latB reduced expression of *YAP1*, whereas expression of endocrine markers—including *INS* and *GCG*—increased (Fig. 4d, Extended Data Figs. 7b, c, 8b). Pancreatic progenitor cells plated on soft hydrogel to reduce stress fibre formation showed enhanced endocrinogenesis in comparison to similar cells plated on a stiff surface (Extended Data Fig. 8e). Finally, western blot analysis showed that latB treatment reduced expression of *HES1*,

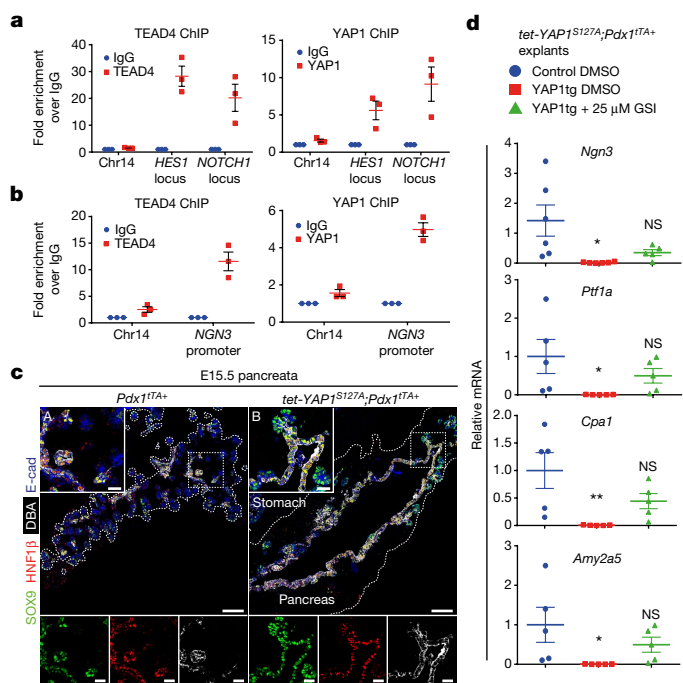


Fig. 3 | Abnormal tip-trunk patterning and impaired differentiation in *Yap1* transgenic pancreas. **a**, ChIP-qPCR of human *HES1* and *NOTCH1* enhancers bound to TEAD4 and YAP1. $n = 3$; data are mean \pm s.e.m. **b**, ChIP-qPCR of a region 250 bp upstream of the ATG start site of the human *NGN3* locus bound to TEAD4 and YAP1. Data represent fold enrichment over pull-down with unrelated IgG. *Chr14* is used as genomic negative control. $n = 3$; data are mean \pm s.e.m. **c**, Co-immunofluorescence analysis for SOX9 (green), HNF1 β (red), DBA (white) and E-cadherin (Ecad, blue) of pancreata from E15.5 control (a) or YAP1tg (b) littermates. Insets show the outlined region at higher magnification and with channels separated for clarity below. The pancreatic epithelium is demarcated by dashed lines. Ducts, marked by SOX9, HNF1 β and DBA, are expanded in the YAP1tg (b) pancreas compared to controls (a). Scale bars, 100 μ m (main panels), 25 μ m (insets). Representative images from three experiments shown. **d**, qRT-PCR analysis of E11.5 explants treated for a further 4 days with either DMSO (controls and YAP1tg) or 25 μ M γ -secretase inhibitor (GSI) (YAP1tg). For *Ngn3*: DMSO (control), $n = 6$; DMSO (YAP1tg), $n = 6$; GSI (YAP1tg), $n = 5$; *Cpa1*: DMSO (control), $n = 5$; DMSO (YAP1tg), $n = 5$; GSI (YAP1tg), $n = 5$; *Ptf1a*: DMSO (control), $n = 5$; DMSO (YAP1tg), $n = 5$; GSI (YAP1tg), $n = 5$; *Amy2a5*: DMSO (control), $n = 5$; DMSO (YAP1tg), $n = 5$; GSI (YAP1tg), $n = 5$. * $P \leq 0.05$, ** $P \leq 0.01$; data are mean expression \pm s.e.m.

suggesting that F-actin-mediated regulation of endocrinogenesis via YAP1 involves Notch signalling (Extended Data Fig. 7d).

Next, we studied whether ECM proteins differentially affect pancreatic cell fate by influencing cell spreading. Although the ability of laminin to inhibit spreading resulted in a higher proportion of confined YAP1[−]NGN3⁺ cells compared to fibronectin (Fig. 4e, Extended Data Fig. 9e, f), spread cells on either fibronectin or laminin failed to induce expression of NGN3 (Extended Data Fig. 9e, f). Furthermore, vitronectin and collagen recapitulated the effects observed with fibronectin and laminin, respectively (Extended Data Fig. 9a–e, g). In vivo, deposition of laminin—but not of fibronectin—in the central epithelium increased at E13.5 and E15.5 compared to E11.5. Whereas fibronectin is mainly deposited around the lumen on the basal side of the epithelium, laminin is distributed on the basal side of the epithelium and as speckles within the epithelium. This means that bi-PPs are more likely to encounter laminin than fibronectin at the secondary transition (Extended Data Figs. 8f, 9i). Consistently, the total number of NGN3⁺ cells in contact with laminin (~69%) was higher than the number of NGN3⁺ cells in contact with fibronectin (~58%) (Extended Data Fig. 9h). The change in the number of cells that contact laminin during the progression to

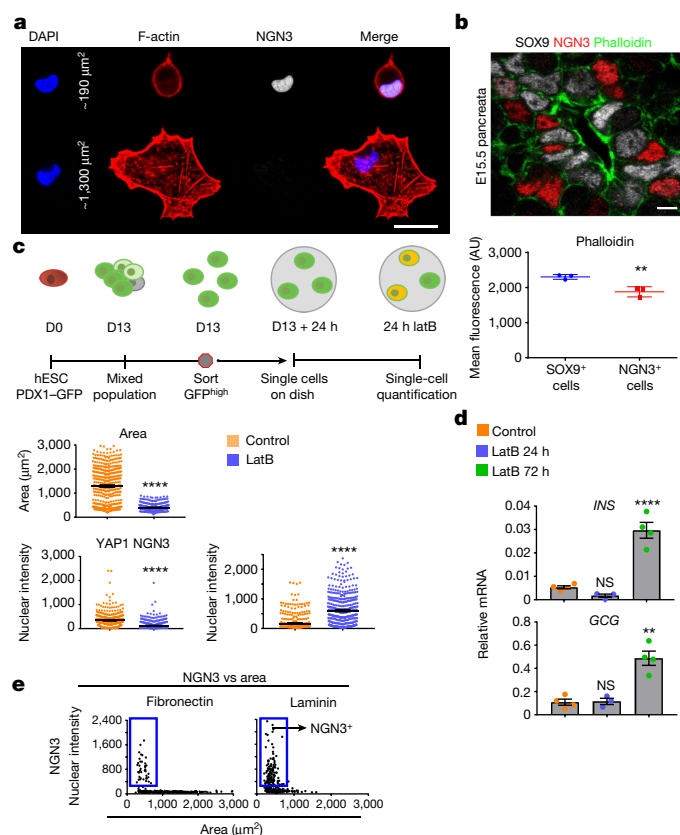


Fig. 4 | Actin dynamics as upstream regulators of endocrinogenesis. **a**, Single pancreatic progenitor cells prepared and sorted as in Fig. 1a and stained with DAPI (blue) and for F-actin (phalloidin, red) and NGN3 (grey). The confined (NGN3⁺) cell (top) has few stress fibres, whereas the spread (NGN3⁻) cell (bottom) has abundant stress fibres. Scale bar, 20 μm; representative images of three independent experiments are shown. **b**, Immunofluorescence analysis for SOX9 (white), NGN3 (red) and F-actin (phalloidin, green) on sections of E15.5 mouse pancreas. Scale bar, 5 μm. Phalloidin mean fluorescence intensity (arbitrary units) was determined for SOX9⁺ or NGN3⁺ cells in 10-μm sections of E15.5 wild-type pancreas using Fiji. *n* = 100 cells from each of three independent pancreata were quantified for phalloidin quantification for each population (SOX9⁺ and NGN3⁺). ***P* = 0.0099 by two-tailed unpaired *t*-test; data are mean ± s.d. **c**, PDX1-GFP⁺ pancreatic progenitor cells sorted on a dish at a single-cell density treated with 1 μM F-actin inhibitor latB or control (DMSO). Single cells were stained for YAP1 and NGN3 and quantified as in Fig. 1a. Each data point representing area, YAP1 signal intensity or NGN3 signal intensity corresponds to an individual cell. *n* = 436 cells (control); *n* = 452 cells (latB); *n* = 3 experiments; *****P* ≤ 0.0001. **d**, Gene expression analysis for insulin (*Ins*) and glucagon (*Gcg*) in unsorted pancreatic precursor culture treated with DMSO or latB on day 13 for 24 or 72 h. Each data point corresponds to one experiment. *n* = 4 independent experiments. ***P* ≤ 0.01, *****P* ≤ 0.0001; ordinary one-way ANOVA; data are shown as mean ± s.e.m. **e**, PDX1-GFP⁺ pancreatic progenitors sorted at single-cell density on dishes coated with fibronectin or laminin, fixed after 24 h and stained for NGN3. Each data point corresponds to an individual cell. Fibronectin, *n* = 284; laminin, *n* = 286; *n* = 3 independent experiments. Blue boxes indicate the NGN3⁺ cell population.

the endocrine precursor stage suggests that laminin acts as an inducer of endocrine differentiation. Together with the observation that inhibition of focal adhesion kinase (FAK) activity reduces YAP1 activity and promotes endocrine specification²⁹ (Extended Data Figs. 7b–d, 8c, d), these findings suggest that ECM-mediated induction of endocrinogenesis via YAP1 is mediated by reduced integrin–FAK signalling.

Next, we screened for integrin α subunits that were expressed in pancreatic progenitors and expressed at a lower level in NGN3⁺ endocrine precursors. Of the tested integrins, only integrin α5 exhibited

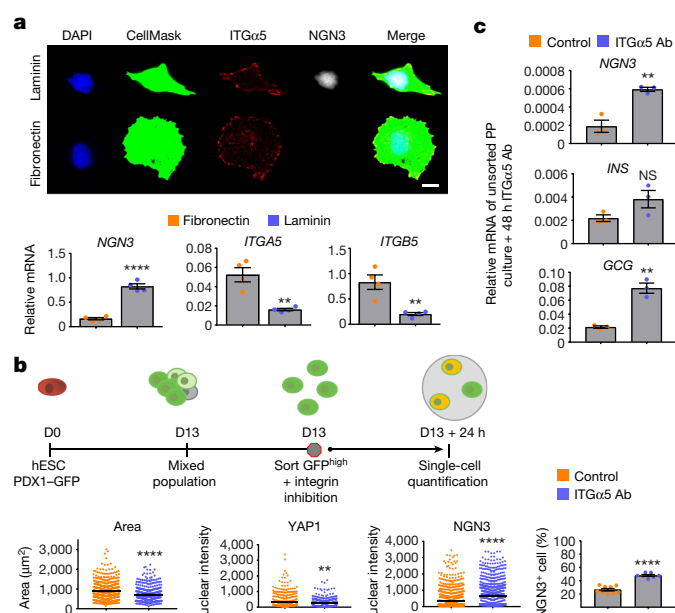


Fig. 5 | Fibronectin–integrin α5β1–YAP1 signalling axis inhibits endocrinogenesis. **a**, Pancreatic progenitor cells prepared and sorted as in Fig. 4c and analysed after 24-h plating on fibronectin or laminin. Top, immunofluorescence for integrin α5 (red) and NGN3 (grey). DAPI, blue; CellMask, green. Scale bar, 10 μm. Bottom, expression of NGN3, *ITGA5* and *ITGB1*. *n* = 4; ***P* ≤ 0.01, *****P* ≤ 0.0001; two-tailed unpaired *t*-test; data are mean ± s.e.m. **b**, Pancreatic progenitors sorted on a dish at a single-cell density treated with function-blocking integrin α5 antibody or isotype antibody (control) for 24 h. Single cells stained for YAP1 and NGN3 and quantified as in Fig. 1a. Each data point representing area (isotype, *n* = 549; integrin α5, *n* = 426), YAP1 (isotype, *n* = 549; integrin α5, *n* = 426), and NGN3 corresponds to an individual cell (isotype, *n* = 936; integrin α5, *n* = 904). Data aggregated from 4 independent experiments; ***P* ≤ 0.01, *****P* ≤ 0.0001. Right, percentage of NGN3⁺ cells was calculated as in Extended Data Fig. 9f. Each data point representing NGN3⁺ cells corresponds to one independent experiment. **c**, Unsorted hESC-derived pancreatic progenitor cell culture treated with control or inhibiting integrin α5 antibody for 48 h. Expression of endocrine genes NGN3, *INS* and *GCG*. ***P* ≤ 0.01 by two-tailed unpaired *t*-test; data are shown as mean expression ± s.e.m.; 3 independent experiments.

lower expression in isolated NGN3⁺ cells (Extended Data Fig. 9j–n). Notably, plating cells on laminin reduced expression of integrin α5, total FAK, phosphorylated FAK (pFAK) and YAP1, and increased that of NGN3 (Fig. 5a, Extended Data Fig. 10a). These data are consistent with the role of integrin α5β1 as a canonical fibronectin receptor that promotes cell spreading³⁰. Most importantly, these data suggest that exposure to certain ECM proteins, such as laminin and collagen, promotes endocrine specification via reduced expression of integrin α5 in pancreatic progenitors. Indeed, treating single pancreatic progenitors with a function-blocking antibody against integrin α5 reduced spreading and expression of pFAK and YAP1, whereas NGN3 expression and the number of NGN3⁺ cells increased (Fig. 5b, Extended Data Fig. 10b). Using unsorted hESC-derived pancreatic progenitors, we extend this data by showing that treatment with an inhibitory antibody against integrin α5 or an siRNA against *ITGA5* not only results in enhanced expression of NGN3 (also validated by increased NGN3 promoter activity), but also enhances expression of mature endocrine markers, including *INS* and *GCG* (Fig. 5c, Extended Data Fig. 10c). To verify the relevance of these results in vivo, we found that at E14.5 NGN3⁺ endocrine precursors display diminished integrin α5 expression compared to nearby cells in the trunk (Extended Data Fig. 10d), and that *Itga5* and *Yap1* mRNA expression is reduced in isolated E15.5 endocrine cells compared to isolated bi-PPs (Extended Data Fig. 10e, f). Furthermore, inhibiting integrin–fibronectin interactions in the human

fetal pancreas (using arginylglycylaspartic acid (RGD) peptides) increased the number of endocrine cells³¹, suggesting that collagen- and laminin-binding integrins stimulate endocrinogenesis *in vivo*.

Our results provide new insights into how changes in integrin $\alpha 5 \beta 1$ expression cell-autonomously determine YAP1-mediated fate choices in bi-PPs (Extended Data Fig. 10g). ECM deposition and positions of cells are in continuous flux *in vivo*, implying a dynamic ECM exposure at the single-cell level. Therefore, we do not consider cell spreading *per se*, but rather the corresponding mechanical influences mediated by integrin $\alpha 5 \beta 1$ -triggered actin cytoskeletal remodelling, to be relevant to cell fate decisions *in vivo*. This scenario suggests that the ultimate fate of each bi-PP during pancreas development is governed by its unique history of ECM exposure, which ultimately controls the expression level of integrin $\alpha 5 \beta 1$. Bi-PPs that encounter a laminin- or collagen-enriched milieu fail to maintain integrin $\alpha 5 \beta 1$ expression, leading to activation of *NGN3* (reduced repression by YAP1–TEAD4–HES1) and eventually endocrine differentiation. By contrast, bi-PPs that are exposed to fibronectin- or vitronectin-enriched ECM maintain integrin $\alpha 5 \beta 1$ -expression and fail to activate the endocrine program (*NGN3* repression is maintained by YAP1–TEAD4–HES1^{9,10}). As a consequence, bi-PPs maintain their progenitor state, which eventually leads to the default commitment to the duct lineage³². Furthermore, we identify YAP1 as the main transcriptional gatekeeper that responds to integrin $\alpha 5 \beta 1$ -mediated cellular tension via F-actin bundling.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0762-2>.

Received: 24 November 2017; Accepted: 19 October 2018;

Published online 28 November 2018.

- Shih, H. P., Wang, A. & Sander, M. Pancreas organogenesis: from lineage determination to morphogenesis. *Annu. Rev. Cell Dev. Biol.* **29**, 81–105 (2013).
- Pan, F. C. & Wright, C. Pancreas organogenesis: from bud to plexus to gland. *Dev. Dyn.* **240**, 530–565 (2011).
- Ameri, J. et al. Efficient generation of glucose-responsive beta cells from isolated GP2⁺ human pancreatic progenitors. *Cell Reports* **19**, 36–49 (2017).
- Rezania, A. et al. Maturation of human embryonic stem cell-derived pancreatic progenitors into functional islets capable of treating pre-existing diabetes in mice. *Diabetes* **61**, 2016–2029 (2012).
- Kilian, K. A., Bugarija, B., Lahn, B. T. & Mrksich, M. Geometric cues for directing the differentiation of mesenchymal stem cells. *Proc. Natl Acad. Sci. USA* **107**, 4872–4877 (2010).
- Mosqueira, D. et al. Hippo pathway effectors control cardiac progenitor cell fate by acting as dynamic sensors of substrate mechanics and nanostructure. *ACS Nano* **8**, 2033–2047 (2014).
- Warmflash, A., Sorre, B., Etoc, F., Siggia, E. D. & Brivanlou, A. H. A method to recapitulate early embryonic spatial patterning in human embryonic stem cells. *Nat. Methods* **11**, 847–854 (2014).
- Dupont, S. et al. Role of YAP/TAZ in mechanotransduction. *Nature* **474**, 179–183 (2011).
- Cebola, I. et al. TEAD and YAP regulate the enhancer network of human embryonic pancreatic progenitors. *Nat. Cell Biol.* **17**, 615–626 (2015).
- George, N. M., Day, C. E., Boerner, B. P., Johnson, R. L. & Sarvetnick, N. E. Hippo signaling regulates pancreas development through inactivation of Yap. *Mol. Cell Biol.* **32**, 5116–5128 (2012).
- Hingorani, S. R. et al. Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse. *Cancer Cell* **4**, 437–450 (2003).
- Zhang, N. et al. The Merlin/NF2 tumor suppressor functions through the YAP oncoprotein to regulate tissue homeostasis in mammals. *Dev. Cell* **19**, 27–38 (2010).
- Kopp, J. L. et al. Sox9⁺ ductal cells are multipotent progenitors throughout development but do not produce new endocrine cells in the normal or injured adult pancreas. *Development* **138**, 653–665 (2011).
- Liu-Chittenden, Y. et al. Genetic and pharmacological disruption of the TEAD–YAP complex suppresses the oncogenic activity of YAP. *Genes Dev.* **26**, 1300–1305 (2012).
- Gao, T. et al. Hippo signaling regulates differentiation and maintenance in the exocrine pancreas. *Gastroenterology* **144**, 1543–1553 (2013).
- Jansson, L. & Larsson, J. Normal hematopoietic stem cell function in mice with enforced expression of the Hippo signaling effector YAP1. *PLoS ONE* **7**, e32013 (2012).
- Holland, A. M. & Hale, M. a, Kagami, H., Hammer, R. E. & MacDonald, R. J. Experimental control of pancreatic development and maintenance. *Proc. Natl Acad. Sci. USA* **99**, 12236–12241 (2002).
- Zhao, B. L. L. and K.-L. G. Hippo signaling at a glance. *J. Cell Sci.* **126**, 2135–2140 (2010).
- Murtaugh, L. C., Stanger, B. Z., Kwan, K. M. & Melton, D. A. Notch signaling controls multiple steps of pancreatic differentiation. *Proc. Natl Acad. Sci. USA* **100**, 14920–14925 (2003).
- Hald, J. et al. Activated Notch1 prevents differentiation of pancreatic acinar cells and attenuate endocrine development. *Dev. Biol.* **260**, 426–437 (2003).
- Totaro, A. et al. YAP/TAZ link cell mechanics to Notch signalling to control epidermal stem cell fate. *Nat. Commun.* **8**, 15206 (2017).
- Lee, J. C. et al. Regulation of the pancreatic pro-endocrine gene *Neurogenin3*. *Diabetes* **50**, 928–936 (2001).
- Beyer, T. A. et al. Switch enhancers interpret TGF- β and Hippo signaling to control cell fate in human embryonic stem cells. *Cell Reports* **5**, 1611–1624 (2013).
- Kim, M., Kim, T., Johnson, R. L. & Lim, D.-S. Transcriptional co-repressor function of the Hippo pathway transducers YAP and TAZ. *Cell Reports* **11**, 270–282 (2015).
- Horn, S. et al. *Mind bomb 1* is required for pancreatic β -cell formation. *Proc. Natl Acad. Sci. USA* **109**, 7356–7361 (2012).
- Halder, G., Dupont, S. & Piccolo, S. Transduction of mechanical and cytoskeletal cues by YAP and TAZ. *Nat. Rev. Mol. Cell Biol.* **13**, 591–600 (2012).
- Zhao, B. et al. Cell detachment activates the Hippo pathway via cytoskeleton reorganization to induce anoikis. *Genes Dev.* **26**, 54–68 (2012).
- Tian, B., Gabelt, B. T., Kaufman, P. L. & Geiger, B. in *Encyclopedia of the Eye* (ed. Dartt, D. A.) 549–555 (Academic, Cambridge, 2010).
- Afrikanova, I. et al. Inhibitors of Src and focal adhesion kinase promote endocrine specification: impact on the derivation of β -cells from human pluripotent stem cells. *J. Biol. Chem.* **286**, 36042–36052 (2011).
- Huveners, S. & Danen, E. H. J. Adhesion signaling—crosstalk between integrins, Src and Rho. *J. Cell Sci.* **122**, 1059–1069 (2009).
- Cirulli, V. et al. Expression and function of $\alpha v \beta 3$ and $\alpha v \beta 5$ integrins in the developing pancreas: roles in the adhesion and migration of putative endocrine progenitor cells. *J. Cell Biol.* **150**, 1445–1459 (2000).
- Zhou, Q. et al. A multipotent progenitor domain guides pancreatic organogenesis. *Dev. Cell* **13**, 103–114 (2007).

Acknowledgements We thank D. Pan for Yap1 floxed mice, L. Jansson and J. Larsson for Tet-O-Yap1 mice, Y. H. Kim and A. Grapin-Botton for *Ngn3*-RFP mice, Beta Cell Biology Consortium (1 U01 DK089570-01) for supplying antibodies, J. P. Larsen and S. E. Christine for assistance with human pluripotent stem cell expansion and differentiation experiments, G. de la Cruz and P. van Dieken (DanStem FACS core facility), J. Bulkescher (DanStem Imaging core facility), and G. Karemore for assistance with image analysis and statistics. We thank D. Kluver Hansen and A. Stiehm for technical assistance. A.M. is a recipient of a post-doctoral fellowship from Lundbeck Foundation (R151-2013-14359). H.S. and P.S. are recipients of grant HumEn project funded by the European Commission's Seventh Framework Programme for Research (agreement 602587). P.S. received a grant from Novo Nordisk Foundation (NNF10717), H.S. also received grants from the Lundbeck Foundation (R100-A9422) and the Danish Council for Independent Research (ID: DFF—1331-00310A), Danish Council for Strategic Research. The Novo Nordisk Foundation Center for Stem Cell Biology is supported by a Novo Nordisk Foundation grant number NNF17CC0027852.

Reviewer information Nature thanks F. Spagnoli and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions A.M., C.P. and H.S. conceived and designed the experiments and wrote the manuscript. P.A.S. performed immunofluorescence staining in Fig. 3c, Extended Data Figs. 3b, 5b, *in vivo* actin quantifications in Fig. 4b, and Notch mutant analysis in Extended Data Fig. 5e, f. K.H.d.L. performed ChIP experiments in Fig. 3b, Extended Data Fig. 6a and gene expression in Extended Data Fig. 10e, f. A.J. performed staining in Extended Data Fig. 4h. P.S. contributed to materials and data analysis.

Competing interests A.M. and H.S. are named as inventors in international patent application (WO2016170067A1), which is based on this work. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0762-2>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0762-2>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to H.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Mice. Mice were housed at the University of Copenhagen and all experiments were performed according to guidelines and ethics approved by the Danish Animal Experiments Inspectorate (Dyreforsøgstilsynet). Pdx1-tTA¹⁷, tetO-Yap1¹⁶, Pdx1-Cre¹¹, Sox9-CreER^{T213,33}, Yap1^{fl/fl12}, Foxa2^{T2AiCre25}, Rosa26^{dnMaml1-eGFP25}, Ngn3-trFP³⁴ and Hes1-eGFP³⁵ mouse strains were used. Data were collected from both male and female embryos.

Immunoblotting. Explants, embryonic pancreata or differentiated hESC were lysed in RIPA buffer containing 1 × phosphatase inhibitor cocktail (Sigma P5726) and 1 × protease inhibitor (Thermo Scientific). All cellular material was sonicated at 4 °C and spun down at full speed for 5 min on a tabletop centrifuge at 4 °C. Samples were boiled in sample loading buffer (Novex) and separated by SDS-PAGE (Novex) before transfer onto a nitrocellulose membrane (GE Healthcare). Membranes were blocked for 1 h in 1% bovine serum albumin (BSA) in TBS-tween (Novex) and incubated with primary antibodies overnight at 4 °C in blocking solution. Blots were washed and incubated with HRP-conjugated secondary antibodies for 1 h and proteins were visualized by chemiluminescence (GE Healthcare). Relative band intensities were analysed using Fiji and normalized to housekeeping genes. Primary antibodies used for blotting are as follows: YAP1 (CS-4912), HES1 (CS-11988), FAK (CS-13009), pFAK Y397 (CS-8556), GAPDH (ab-8245), PDX1 (BCBC-2028), integrin $\alpha 5$ (ab150361), α -tubulin (Sigma-9026-clone DM1A) and vinculin (Sigma-V9131).

Blood glucose measurements. Random fed blood glucose was measured using a hand-held glucometer (OneTouch Ultra; Lifescan). Blood was obtained from the pups after they were killed.

Sorting E15.5 bipotent and endocrine progenitors for gene expression. Pools of 5 E15.5 pancreata from Hes1-eGFP;Ngn3-trFP mice were dissected in PBS, dissociated using Liberase TL (117 U/ml; Roche) with shaking for 20 min at 37 °C and subsequently with 0.125% Trypsin-EDTA (Gibco), for 10 min at 37 °C, both with DNase I (300 U/ μ l; Invitrogen). Dissociation was stopped with 10% FBS in PBS. Dissociated cells were incubated with 1:200 DBA-biotin (Vector Labs) for 10 min at 4 °C, washed in fetal bovine serum (FBS) in PBS, and incubated with 1:1,000 streptavidin-APC-Alexa Fluor 750 for 15 min at 4 °C. Cells were sorted on a Sony SH800 directly into RLT buffer for immediate RNA purification using a Qiagen microElute RNA purification kit according to the manufacturer's instructions. Quality and quantity of RNA was checked using a Bioanalyser Pico kit (Agilent) and 25 ng RNA from each sample was loaded on Agilent SurePrint 8x60k Mouse microarrays according to the manufacturer's instructions. Results were analysed for differential expression using the Limma package³⁶.

Cell culture, differentiation and FACS sorting of pancreatic progenitor cells for re-culture and analysis. Undifferentiated PDX1-EGFP hESC³ reporter cells were maintained in DEF-CS culture system (Takara) and passaged with TrypLE (Life Technologies)³. All hESC-derived pancreatic progenitors (hESC-PP) were differentiated using PDX1-EGFP hESC reporters unless mentioned otherwise. At 80–90% confluency, undifferentiated hESCs were differentiated into pancreatic endoderm (PE) stage following a modified Rezanian 2D protocol^{13,37}. Differentiated cells were dissociated at PE stage with Accutase for 15–20 min at 37 °C and washed with FACS buffer (EDTA 3 μ M, BSA 0.1%, HBSS (Gibco 14185-052)). Cell sorting was performed using a Sony SH800. EGFP flow cytometer calibration beads (Clontech) were used as a reference to gate for GFP^{high}, GFP^{low} and GFP[−] populations. The beads gave six distinct fluorescence intensity peaks and provided consistent gating and sorting of GFP^{high} pancreatic progenitor cells between different experiments. For NKX6.1 single-cell analysis, GP2 markers were used for sorting as previously described³. 7AAD or DAPI were used as a live/dead cell markers. For re-culturing of single pancreatic progenitors, cells were sorted into post-sort medium (PE stage medium + 10 μ M Y27632 + penicillin–streptomycin), centrifuged, re-stained with post-sort medium and plated at low density. The NGN3–GFP reporter line³⁸ was differentiated into pancreatic endoderm stage following the Rezanian protocol on Matrigel-coated dishes³⁹ until S5 stage and dissociated with TrypLE and resuspended in PBS + 0.5% BSA. Cells were sorted by FACS to obtain fractions of GFP^{high}, GFP^{low} and GFP[−] populations using the same gating for all experiments. Antibodies used for FACS analysis in Extended Data Fig. 9m are Biolegend NK1-SAM1 (328002) mouse anti-integrin $\alpha 5$ and Biolegend purified mouse IgG2b, k isotype control antibody (400302). All the cell lines tested negative for mycoplasma. **Cell treatment and transfections.** Small molecule inhibition was performed using verteporfin (Sigma SML0534; 1 μ g/ml), latrunculin B (Sigma; 0.5–1 μ M) and PF-573228 FAK inhibitor (Sigma; 5 μ M). Integrin $\alpha 5$ inhibition assays were performed using function-blocking NK1-SAM-1 CD49e antibody (Biolegend 10 μ g/ml). For single-cell treatment, approximately 2.5×10^4 sorted pancreatic progenitor cells were seeded on fibronectin-coated 35-mm μ -Dish (Corning). Small molecule inhibition started after adhesion of post-sorted cells, whereas integrin $\alpha 5$

function-blocking antibody was incubated with post-sorted cells before adhesion. Treatments lasted 24 h. Unsorted hESC pancreatic progenitor cultures were treated at the PE stage for the indicated period of time (24–72 h).

Unsorted hESC-pancreatic progenitor cultures were transfected with siRNAs at day 13 using Lipofectamine RNAiMAX (Thermo Fisher), and with plasmid DNA with Lipofectamine 2000 (Thermo Fisher), according to the manufacturer's instructions and collected after 72 h for further analysis. A pool of 2 Ambion silencer select validated siRNAs were used for each gene: YAP1 (s20366 and s20368), HES1 (s6921 and s6922) and ITGA5 (s7548 and s7549).

For luciferase assays, hESC-pancreatic progenitor culture was transfected at day 13 with siRNA (first day) and with plasmid DNA (second day), and collected 24 h after DNA transfection. Cells were collected in 1 × PLB buffer (Promega E1910 Dual Luciferase Assay Kit) and luciferase activity was analysed using LUMIstar Omega. Normalization was carried out based on co-transfected Renilla activity. Each experiment was repeated at least three times.

Explant culture and inhibition assays. E11.5 mouse embryonic dorsal pancreata were microdissected and cultured on fibronectin-coated (Life Technologies 33010-018) 24-well plates as previously described^{40,41}. The culture medium contained M199, 10% fetal bovine serum, 1% penicillin–streptomycin and 0.5% Fungizone. Medium was changed every second day. Inhibitors were added according to experiment and the explants were processed for qRT-PCR, whole-mount staining or western blot analysis. Inhibitors used were Gamma Secretase Inhibitor XXI, CompoundE (565790-MERCK; 25 μ M), verteporfin (Sigma SML0534; 1 μ g/ml), latrunculin B (Sigma; 1 μ M) and PF-573228 FAK inhibitor (Sigma; 3 μ M).

NGN3 promoter assay construct design and mutagenesis. A genomic region 800 bp upstream of the human NGN3 coding region was amplified with restriction-site-overhang oligonucleotides based on genomic DNA and cloned in the multiple cloning site region of the pGL3 vector using Kpn1 and Pst1 sites. TEAD and HES1 binding-site mutations of the 800-bp region were inserted using DNA synthesized from IDT and cloned using same restriction sites in the pGL3 vector backbone. Sequences of all constructs were verified, and constructs were used for transient transfections on hESC-pancreatic progenitor day 13 cells in 12 or 24-well plates.

Micropatterned substrate preparation and experimental setup. Micropatterned glass slides were purchased from CYTOO. Disc and square islands of different sizes were printed on the slides with sufficient gap in between islands to ensure no cell–cell contact. ECM coating was performed according to the manufacturer's protocol. In brief, the micropatterned substrates were incubated at room temperature for 2 h or at 4 °C overnight with fibronectin (Sigma F0895; 100 μ g/ml). Seeding of sorted PDX1^{high} pancreatic progenitor cells at 100,000 cells/chip resulted in low (<0.1%) attachment efficiency, but produced a high rate of single-cell attachment. Increasing the number of cells led to doublets or multiple cells in each island, which were excluded from quantification during analysis. Cells were incubated in post-sort medium and fixed with 4% PFA for end-point analysis after 24 h. For experiments on substrate compliance, fibronectin-coated hydrogels (10 kPa) and glass slides ($\geq 2,000,000$ kPa) were prepared by CYTOO. For self-aggregation experiments, low concentration (2%) Matrigel⁴² was added in the post-sort medium to trigger clumping of sorted PDX1^{high} pancreatic progenitor cells.

Immunostaining, imaging and image analysis. Samples were carefully washed and fixed with either 4% PFA or formaldehyde solution for 10–20 min at room temperature. Actin and DNA were stained using phalloidin (Sigma) and DAPI, respectively. Primary antibody incubation was performed overnight at 4 °C with the following antibodies: PDX1 (R&D Systems; 1:500), NKX6.1 (DSHB; 1:200), SOX9 (Millipore; 1:500), NGN3 (R&D Systems; 1:400), FOXA2 (Santa Cruz; 1:500), HNF6 (Santa Cruz; 1:200), YAP1 (Santa Cruz; 1:200), GFP (Abcam; 1:500), insulin (Dako; 1:1,000), glucagon (Linco Research; 1:1,000), biotinylated DBA (Vector Labs, cat # B-1035; 1:500). Secondary antibody incubation was carried out for 60 min at room temperature or overnight at 4 °C. Images were acquired using a confocal microscope with Zeiss LSM 780 $\times 20$ (0.8 NA) and $\times 40$ (1.3 NA), or Leica SP8 20 \times (0.75 NA) and $\times 40$ (1.3 NA) objectives. For fluorescence intensity quantification, Volocity 6.2 (PerkinElmer) or Imaris 9.0.2 (Bitplane) software were used. For single-cell experiments, individual cell areas were segmented using CellMask green stain (ThermoFisher), a nonspecific marker which produced better signal than actin in low-tensioned cells. Nuclei were segmented using the DAPI signal. Only nuclear signal was used for quantifications of transcription factor immunofluorescence, such as for PDX1, NGN3, YAP1, SOX9, FOXA2 and HNF6. Live-cell imaging presented in Extended Data Fig. 2a was performed using a Leica AF6000 Widefield Screening platform $\times 20$ (0.40 NA) in an environmental chamber (5% CO₂, 37 °C). The NGN3⁺ and insulin⁺ fractions in hESC differentiated cultures were measured in Fiji (ImageJ 1.50f) as the number of NGN3⁺/insulin⁺ cells per DAPI⁺ area.

Mouse pancreatic dissection, fixation, embedding, sectioning and immunofluorescence analysis was performed as described⁴⁰. Ten-micrometre sections, mounted in Vectashield with or without DAPI (Vector Labs), were imaged on a

Leica SP8 confocal laser-scanning microscope. The following antibodies and stains were used: rabbit monoclonal antibody against YAP1 (D8H1X) (Cell Signaling Technology 14074; 1:200); guinea-pig anti-Sox9 (BCBC, a gift from O. Madsen, Novo Nordisk; 1:2,000); guinea-pig anti-Pdx1 (BCBC AB2028; 1:1,000); guinea-pig anti-insulin (DAKO A0564; 1:800); rat anti-E-cadherin (Takara M108-clone ECCD-2; 1:400); mouse anti-Nkx6.1 (DSHB F55A12; 1:500); mouse monoclonal anti-glucagon (Sigma Aldrich G2654; 1:500); chicken polyclonal anti-GFP (Abcam ab13970; 1:1,000); rabbit anti-Ngn3 (BCBC AB2011; 1:4,000); rabbit anti-Hnf1 β (Santa Cruz Biotechnology sc-22840; 1:2,000); Armenian hamster anti-Muc1 (Thermo Fisher MA5-11202; 1:200); rat anti-laminin (Acris BM6046P; 1:200); rabbit anti-fibronectin (Abcam ab2413; 1:200); biotinylated DBA (Vector Labs B-1035; 1:500); DAPI (Life Tech D1306; 1:1,000); F-actin was detected with Acti-stain 488/phalloidin (Cytoskeleton, PHDG1; 1:200). All secondary antibodies were donkey-raised (Jackson Immuno): DyLight 405 was used at 1:200; Cy3 and A488 were used at 1:1,000 and Cy5 was used at 1:500. The ratio of insulin⁺ or NGN3⁺ cells was estimated by counting the number of NGN3⁺ or insulin⁺ cells per mm² of E-cadherin⁺ pancreas area or DAPI⁺ pancreas area on 8- μ m sections from the embryos or from whole-mount immunostained and imaged explant optical sections.

Gene expression analysis. Total RNA was extracted using a RNeasy Micro Kit for small samples and Mini Kit for larger samples (Qiagen) followed by removal of genomic DNA (DNase I Qiagen) and cDNA synthesis using SuperScript III and oligo(dT) (Invitrogen), according to the manufacturer's instructions. Transcript levels were measured using Taqman assays (Life Technologies) or power SYBR green (Applied Biosystems 4367659) on a StepOnePlus system (Applied Biosystems). For all hESC samples, relative mRNA expression was normalized to *GAPDH* expression. For all mouse samples, relative mRNA expression was normalized to the expression of *HPRT*. Data are shown as mean expression \pm s.e.m. When indicated, fold change was shown for the treated sample in comparison to control sample.

Taqman human primers used include: Hs00371734_g1 (YAP1), Hs00371735_m1 (YAP1), Hs01127536_m1 (ITGB1), Hs01547673_m1 (ITGA5), Hs00232764_m1 (FOXA2), Hs00413554_m1 (HNF6), Hs00236830_m1 (PDX1), Hs00165814_m1 (SOX9), Hs00232355_m1 (NKX6.1), Hs00603586_g1 (PTF1A), Hs04260396_g1 (Klf6), Hs00159598_m1 (NEUROD1), Hs00172878_m1 (HES1), Hs00173014_m1 (PAX4), Hs00534343_s1 (MAFB), Hs00357871_s1 (INSM1), Hs00158126_M1 (ISL1), Hs02758991_G1 (GAPDH), Hs01056157_M1 (CPA1), and Hs01875204_S1 (NGN3).

Taqman mouse primers used include: Mm00494236_m1 (Yap1), Mm00446968_m1 (Hprt), Mm00731595_gH (Insulin2), Mm01269055_m1 (Glucagon), Mm00439797_m1 (Integrin α 5), Mm01253230_m1 (Integrin β 1), Mm00437606_s1 (Ngn3), Mm00493507_m1 (Tead1), Mm00449004_m1 (Tead2), Mm00449013_m1 (Tead3), Mm01189836_m1 (Tead4), Mm01289583_m1 (Taz), Mm00513560_m1 (Taz), Mm01250509_g1 (pancreatic polypeptide), Mm00445450_m1 (Ghrelin), Mm00436671_m1 (Somatostatin), Mm00448840_m1 (Sox9), Mm01976556_s1 (FoxA2), Mm00479622_m1 (Ptf1a).

Mouse sequences of SYBR green oligonucleotides used include: *Birc5* forward GAGGCTGGCTTCATCCACTG, reverse CTTTTTGCTTGTGTGGTCTCC; *Cdc20* forward GGCACATTTCGATTGGGAACG, reverse TAGTGGGG AGACCAGAGGATGGAG; *Snai2* forward TGTGTCTGCAAGATCTGTGG, reverse TGGAGAAGGTTTTGGAGCAG; *Ctcf* (also known as *Ccn2*) forward GGGCTCTTCTGCGATTTC, reverse ATCCAGGCAAGTCGACATTGGTA; *Hnf1b* forward CGGCAAAAGAATCCAGCAA, reverse AGACCCCTCGTT GCAAACA; *Hprt* forward AGCCCCAAATGGTTAAGGT, reverse CAAGG GCATATCCAACAACA; *Pdx1* forward CCCAGTTTACAAGCTCGCTG, reverse CTCGGGTTCCGCTGTAAG; *Nkx6-1* forward ACTTGGCAGG ACCAGAGAGA, reverse AGATTTCGGGTCCAGAGGTT; *Sox9* forward CCACGGAACAGACTACATC, reverse CTGCTCAGTTACCGATGTC; *Cpa1* forward GACGAGGAGAAGCAGCAGAT, reverse GATGCCAGTGTCAATCC AGA; *Amy2a5* forward AGGTCATTGATCTGGGTGGT, reverse GACATCTT CTCGCATTCCAC; *Ptf1a* forward CTTCAGGGCACTCTCTTTC, reverse CG ATGTGAGCTGTCTCAGGA; *Insm1* forward TGTCTGTAGCGTACGGGTTGT, reverse AAAGCCAGACTCCAGCAGTTC; *Isl1* forward CGGAGAGACATGA TGGTGGTT, reverse GGCTGATCTATGTCGCTTTC; *Neurod1* forward CCAGGGTTATGAGATCGTCAC, reverse TCGTCTGAGAACTGAGACA; *Hes1* forward ATAGCTCCCGGCATTTCCAAG, reverse GCGCGGTA TTTCCCCAACA.

Co-immunoprecipitation. For endogenous protein-protein interaction studies, hESC-pancreatic progenitor cells were differentiated until day 13 in 75 cm² flasks, washed with 1 \times PBS and lysed by sonication in lysis buffer (25 mM HEPES pH7.8, 400 mM KCL, 5 mM EDTA, 5% glycerol, 0.4% NP-40 supplemented with 1 mM dithiothreitol, protease inhibitor cocktail (Roche) and phosphatase inhibitor cocktail (Sigma)). Protein extracts were diluted eightfold with binding buffer (25 mM HEPES pH7.8, 50 mM KCL, 5% glycerol, 0.05% NP40, 2.5 mM MgCl₂) to bring final KCl concentration to 50 mM and NP40 to 0.05%, and subjected to

antibody-bound protein-A sepharose immunoprecipitation for 4 h at 4°C followed by three washes with binding buffer^{43,44}. Antibodies (2 μ g each) for IgG control and/or HES1 (Cell Signaling 11988), YAP1 (Cell Signaling 4912) were used for co-immunoprecipitations.

Chromatin immunoprecipitation. Day13 differentiated PDX1-GFP hESC cells were collected from a T75cm² flask, washed with PBS and collected in tubes. Cells were fixed with 1% formaldehyde (Rockland KHF001) at room temperature for 10 min. Fixation was stopped by addition of glycine to a final concentration of 0.125 M and incubation for 5 min. Cells were washed twice with 1 \times PBS at room temperature and collected in 1 ml 0.5% SDS lysis buffer (100 mM NaCl, 50 mM Tris-HCl pH8.1, 5 mM EDTA pH8.0, 0.2% Na₃, 0.5% SDS) and diluted 2:1 with Triton buffer (100 mM Tris-HCl pH8.0, 100 mM NaCl, 5 mM EDTA pH 8.0, 0.2% Na₃, 5% Triton X-100). Lysates were sonicated 10 \times 30 s with a Diagenode Bioruptor, centrifuged for 20 min at full speed at 4°C in a benchtop centrifuge and DNA concentration was measured using Q-bit on de-crosslinked DNA⁴⁵. Chromatin (7.5 μ g) was diluted to 1 ml in RIPA buffer with protease inhibitors (Roche), tumbled overnight at 4°C with the following antibodies: YAP1 (ab52771), TEAD4 (sc-101184), HES1 (sc-H-140), rabbit IgG (Cell Signaling), mouse IgG (sc-2025). Samples were incubated with 30 μ l Protein-G Dynabeads (Thermo) for 3 h at 4°C and washed twice in RIPA on a magnet. Precipitates were un-crosslinked and eluted from beads using 1%SDS, 0.1M NaHCO₃ and proteinase K at 65°C with shaking. DNA was purified with Zymo CHIP DNA kit according to the manufacturer's instructions. Quantitative PCR was performed using Applied Biosystems FAST reagents on a Lightcycler 480II according to the manufacturer's instructions using the following primer sets: *NGN3* (120 bp upstream of the ATG start) forward AGCTGGATTCCGGACAAAGG, reverse ATAGGCTAGGACGAAAGCCG; *HES1* (23 kb upstream of start site)⁹ forward GAGTCGCTGACAGACAGTGC, reverse GAGTCGCCTCATTTCTGGTT; *NOTCH1* (26.5 kb upstream of start site)⁹ forward ATTCTTGGGATGCCTGTGTC, reverse GTCTCTGCCTCC TGCTATG; control Chr14(66894932–66895059)⁴⁶ forward GTGGGCCTTTGG ATATCCCT, reverse GACCTTGGCTGTGTGTCTCCT.

Analysis of F/G-actin protein levels. NGN3-GFP hESCs were differentiated in T75 cm² flasks until S5 stage³⁹ and similar numbers of cells (200,000) were sorted by FACS for GFP[−] and GFP^{high} cells in FACS buffer (1 \times PBS, 0.5% BSA), and pelleted at 800g for 3 min. Cells were resuspended in 100 μ l PBS, 0.1% Triton (with protease inhibitors) and incubated for 5 min at 4°C with rocking. Cells were centrifuged at 15,000g at 4°C for 5 min. The soluble portion (supernatant, globular actin) was boiled with NuPAGE 4 \times LDS (Invitrogen NP0007). The Triton-insoluble portion (pellet, predominantly filamentous actin) was resuspended in 100 μ l RIPA buffer and boiled with 4 \times LDS⁴⁷. Equal amounts of GFP[−] and GFP^{high} fractions were separated by SDS-PAGE and western blotted for actin (C4-ab14128). Protein bands were quantified by densitometry using ImageJ software.

Statistical analysis. Statistically significant differences between two or more conditions were analysed by two-tail Student's *t*-test or by multivariate comparison (one-way ANOVA) using GraphPad Prism 7 software. Single-cell experiments were analysed as previously published⁴⁸. Correlation between cell area and nuclear fluorescence intensity were represented with the Spearman correlation coefficient *r*; $|r| \geq 0.7$ indicates strongly correlated, $0.5 \leq |r| < 0.7$ indicates moderately correlated, and $r = 0$ means not related. Correlation significance (*P* value) was calculated when $|r| \geq 0.5$. **P* ≤ 0.05 , ***P* ≤ 0.01 , ****P* ≤ 0.001 , *****P* ≤ 0.0001 . Bar graphs and dot plots were generated by GraphPad PRISM and show mean \pm s.e.m. unless otherwise indicated.

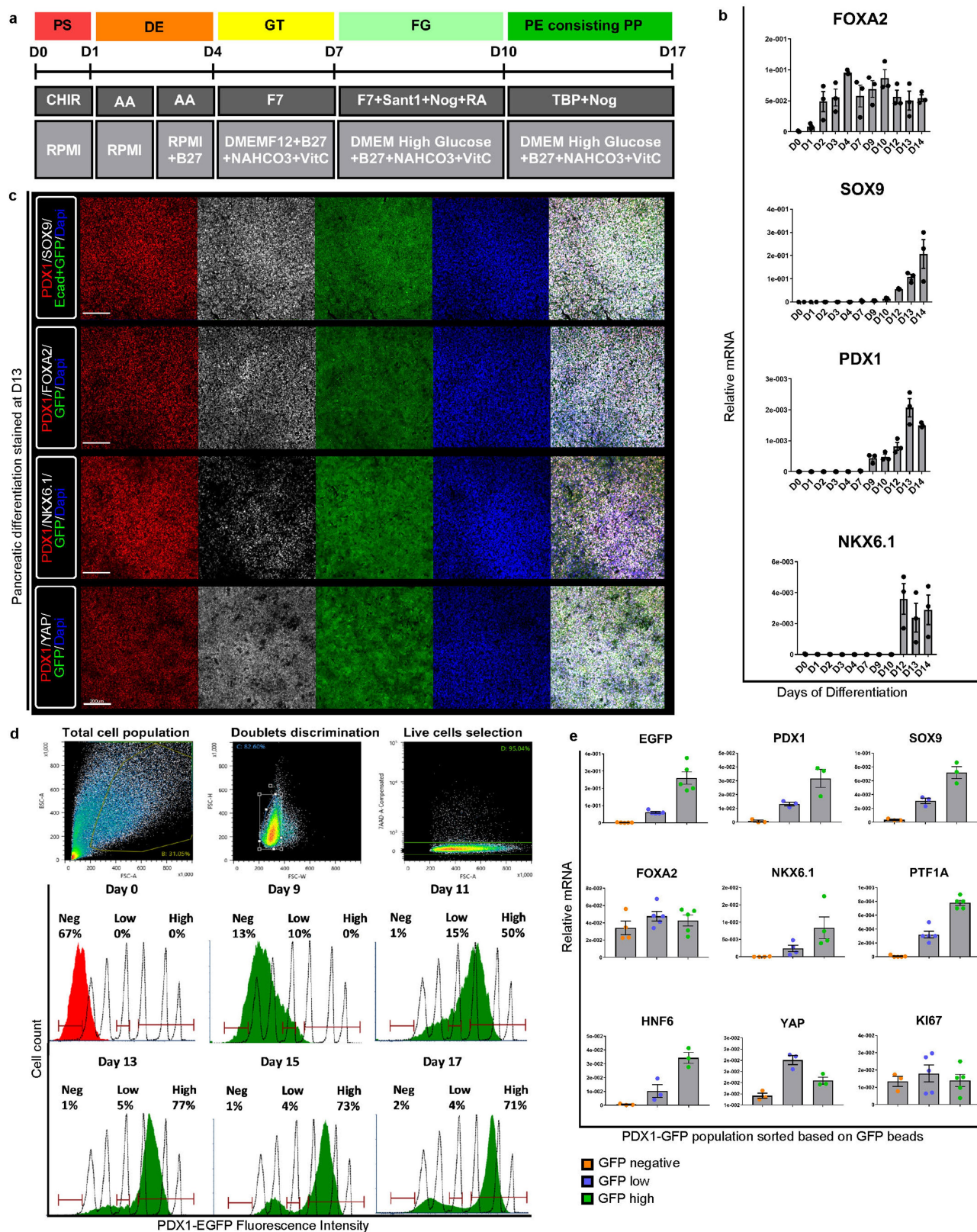
Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The gene expression datasets generated from sorting E15.5 bipotent and endocrine progenitors (Extended Data Fig. 10e, f) have been deposited in the ArrayExpress database under accession code E-MTAB-6891.

- Seymour, P. A. et al. SOX9 is required for maintenance of the pancreatic progenitor cell pool. *Proc. Natl Acad. Sci. USA* **104**, 1865–1870 (2007).
- Kim, Y. H. et al. Cell cycle-dependent differentiation dynamics balances growth and endocrine differentiation in the pancreas. *PLoS Biol.* **13**, e1002111 (2015).
- Klinck, R. et al. A BAC transgenic Hes1-EGFP reporter reveals novel expression domains in mouse embryos. *Gene Expr. Patterns* **11**, 415–426 (2011).
- Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
- Rezanian, A. et al. Enrichment of human embryonic stem cell-derived NKX6.1-expressing pancreatic progenitor cells accelerates the maturation of insulin-secreting cells in vivo. *Stem Cells* **31**, 2432–2442 (2013).
- Löf-Öhlin, Z. M. et al. EGFR signaling controls cellular fate and pancreatic organogenesis by regulating apical polarity. *Nat. Cell Biol.* **19**, 1313–1325 (2017).

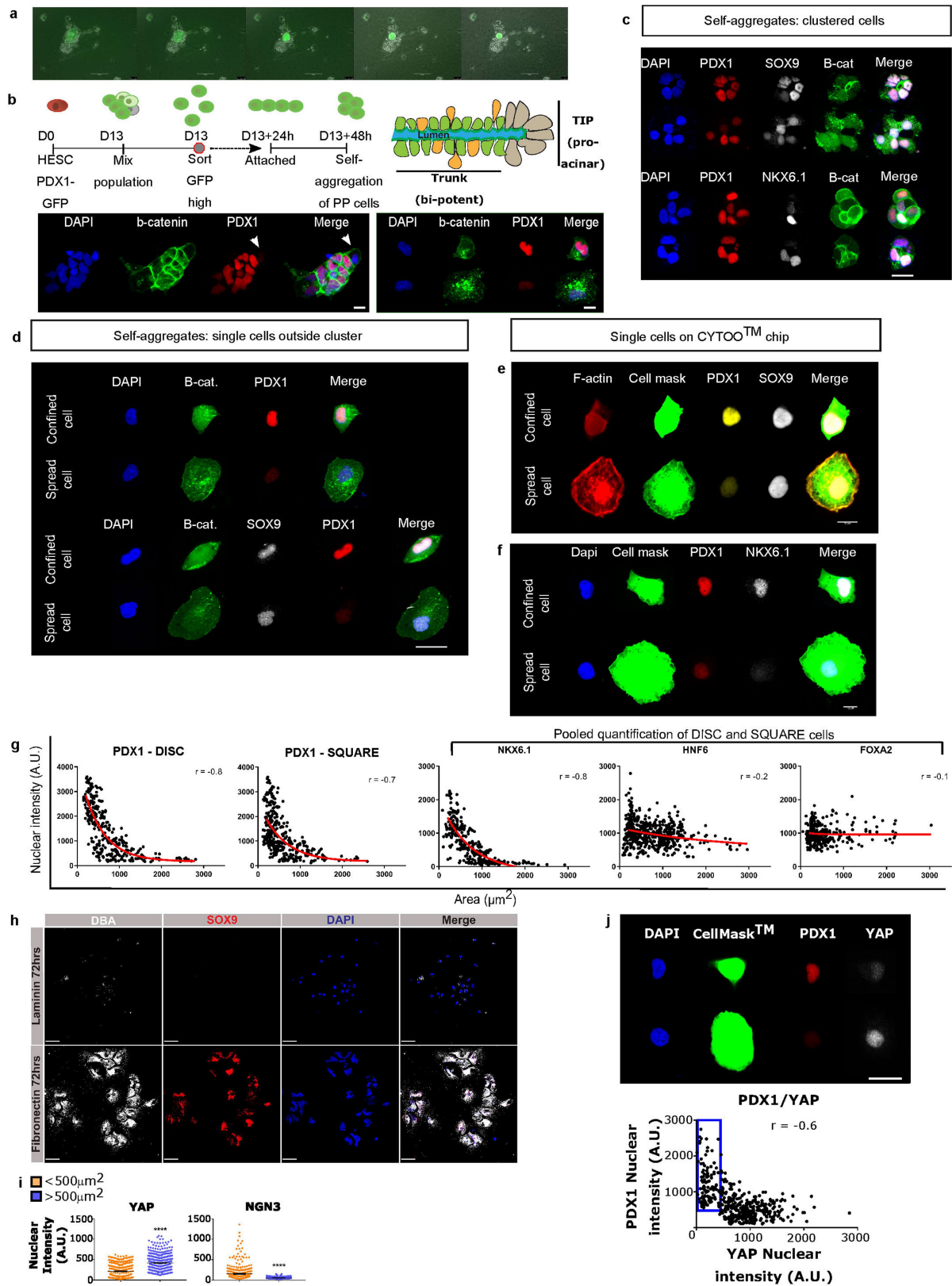
39. Rezaei, A. et al. Reversal of diabetes with insulin-producing cells derived in vitro from human pluripotent stem cells. *Nat. Biotechnol.* **32**, 1121–1133 (2014).
40. Kesavan, G. et al. Cdc42/N-WASP signaling links actin dynamics to pancreatic β cell delamination and differentiation. *Development* **141**, 685–696 (2014).
41. Kesavan, G. et al. Cdc42-mediated tubulogenesis controls cell specification. *Cell* **139**, 791–801 (2009).
42. Rodríguez-Fraticelli, A. E. et al. Developmental regulation of apical endocytosis controls epithelial patterning in vertebrate tubular organs. *Nat. Cell Biol.* **17**, 241–250 (2015).
43. Mamidi, A. et al. Signaling crosstalk between TGF β and Dishevelled/Par1b. *Cell Death Differ.* **19**, 1689–1697 (2012).
44. Dupont, S. et al. FAM/USP9x, a deubiquitinating enzyme essential for TGF β signaling, controls Smad4 monoubiquitination. *Cell* **136**, 123–135 (2009).
45. Funa, N. S. et al. β -catenin regulates primitive streak induction through collaborative interactions with SMAD2/SMAD3 and OCT4. *Cell Stem Cell* **16**, 639–652 (2015).
46. Stein, C. et al. YAP1 exerts its transcriptional control via TEAD-mediated activation of enhancers. *PLoS Genet.* **11**, e1005465 (2015).
47. Parreno, J. et al. Expression of type I collagen and tenascin C is regulated by actin polymerization through MRTF in dedifferentiated chondrocytes. *FEBS Lett.* **588**, 3677–3684 (2014).
48. Schiller, H. B. et al. β_1 - and α_v -class integrins cooperate to regulate myosin II during rigidity sensing of fibronectin-based microenvironments. *Nat. Cell Biol.* **15**, 625–636 (2013).



Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Validation of hESC-derived pancreatic progenitor differentiation. **a**, Schematic of pancreatic progenitor differentiation protocol using the HUES4 hESC PDX1–GFP reporter line. Pancreatic progenitors start to appear from day 10 (D10). **b**, Expression of endodermal and pancreatic genes from day 0 until day 14. *FOXA2* was upregulated early during the definitive endoderm (DE) stage and expression was maintained throughout; this was followed by expression of *PDX1* and *SOX9*, and subsequently *NKX6-1*. Data from two independent experiments are shown as mean expression \pm s.e.m. **c**, Unsorted pancreatic progenitor culture at day 13 was stained for PDX1 (red), DAPI (blue), GFP (green) and either *SOX9*, *FOXA2*, *NKX6.1* or *YAP1* (grey). Representative images from 4 independent experiments are shown. Scale bar, 200 μ m. **d**, FACS analysis of GFP signal shows transition from GFP[−]

to GFP^{low} to GFP^{high} between days 0 and 17. The red bar represents gating of GFP[−] (left), GFP^{low} (middle) and GFP^{high} (right). Gating consistency was ensured by the use of flow cytometer calibration beads for EGFP. GFP^{high} percentage reached saturation after day 13. Representative FACS plots from 4 independent experiments are shown. **e**, At day 13, the three populations in **d** were sorted and analysed by qRT–PCR for gene expression relative to *GAPDH*. GFP^{high} cells were enriched for expression of pancreatic progenitor genes (*PDX1*, *SOX9*, *NKX6-1*, *PTF1A* and *HNF6* (also known as *ONECUT1*)). There was no change in expression of *YAP1* or *MKI67*. EGFP ($n = 5$), *PDX1* ($n = 3$), *SOX9* ($n = 3$), *FOXA2* ($n = 4$), *NKX6-1* ($n = 4$), *PTF1A* ($n = 5$), *HNF6* ($n = 3$), *YAP1* ($n = 2$), *MKI67* ($n = 3$). Data are mean expression \pm s.e.m.

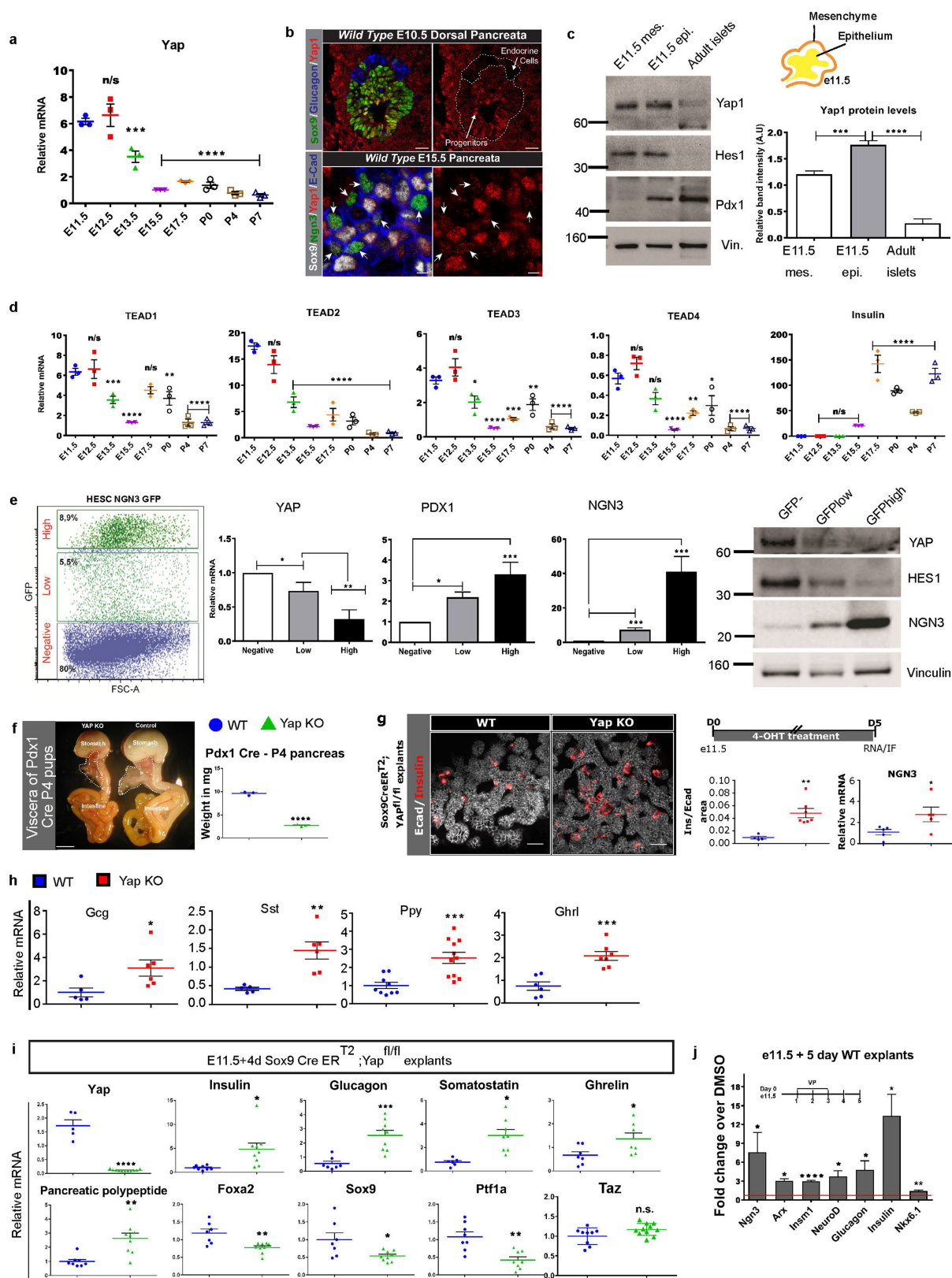


Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Confinement of single pancreatic progenitor cells is associated with YAP1 downregulation, PDX1 maintenance and endocrine commitment.

a, Time lapse of self-aggregation after adhesion of sorted pancreatic progenitors derived from hESCs. GFP represents PDX1 expression. Representative results from three independent experiments are shown. Scale bar, 100 μm . **b**, Self-aggregation of sorted human in vitro pancreatic progenitors after 48 h analysed by immunostaining for β -catenin (green) and PDX1 (red), and DAPI staining (blue). Images are representative of three independent experiments. Top left, schematic of self-aggregation experiments to obtain bi-PPs. Top right, illustration of in vivo bi-PPs within trunk epithelium. Bottom left, cluster formed after self-aggregation of sorted bi-PPs. Scale bar, 10 μm . Bottom right, representation of non-spread (PDX1^{high}) and spread (PDX1^{low}) single cells outside the cluster lacking cell–cell contact. Scale bar, 10 μm . **c**, Cluster from sorted self-aggregated pancreatic progenitor cells stained for pancreatic progenitor markers PDX1 (red), SOX9 or NKX6.1 (grey). β -catenin (green, cell membrane marker), and DAPI (blue). Scale bar, 20 μm . Images are representative of three independent experiments. **d**, Single cells that are not incorporated into clusters during self-aggregation experiments. Representation of PDX1^{high} confined cells (top row of each pair) and PDX1^{low} spread cells (bottom row of each pair) stained for PDX1 (red), SOX9 (grey), β -catenin (green) and DAPI (blue). Scale bar, 20 μm . Images representative of three independent experiments. **e**, Sorted single pancreatic progenitor cells adhered on CYTOO chip, stained for F-actin (red) to mark stress fibres and with CellMask (green) to mask cytoplasm, and immunofluorescence for PDX1 (yellow) and SOX9 (grey). Scale bar, 10 μm . Images representative of three independent experiments. **f**, As in **e**, but stained for PDX1 (red), NKX6.1 (grey), with CellMask (green),

and DAPI (blue). Scale bar, 10 μm . Confined cell (top) and spread cell (bottom). Images representative of three independent experiments. **g**, Single pancreatic progenitor cells on CYTOO chip 24 h after sorting were analysed by immunostaining and quantified as in Fig. 1a. Nuclear intensity of each pancreatic progenitor marker was quantified and plotted against cell area. Each data point corresponds to an individual cell (PDX1-DISC, $n = 309$; PDX1-SQUARE, $n = 366$; NKX6.1, $n = 322$; HNF6, $n = 561$; FOXA2, $n = 344$). Data aggregated from three independent experiments. Spearman's correlation coefficient (r) is calculated for cell area versus nuclear intensity of staining. PDX1, $r = -0.8$ (**** $P \leq 0.0001$); NKX6.1, $r = -0.8$ (**** $P \leq 0.0001$); FOXA2, $r = -0.1$; HNF6, $r = -0.2$. **h**, Co-immunofluorescence analysis of 500,000 sorted PDX1-GFP^{high} hESC-derived pancreatic progenitor cells, plated and cultured for 72 h on 2-well chamber slides, coated with either fibronectin or laminin, in pancreatic progenitor medium. Cells were immunostained for the ductal markers DBA (white) and SOX9 (red), with DAPI staining (blue). Scale bar, 35 μm . Representative images from three independent experiments are shown. **i**, Nuclear intensities of YAP1 and NGN3 in individual cells with low spreading ($<500 \mu\text{m}^2$) and high spreading ($>500 \mu\text{m}^2$). $<500 \mu\text{m}^2$, $n = 365$; $>500 \mu\text{m}^2$, $n = 541$; 3 independent experiments. Data are shown as mean \pm s.e.m.; two-tailed unpaired t -test. **j**, Expression correlation (r) between PDX1 and YAP1. Top, staining for DAPI (blue), PDX1 (red), YAP1 (grey) and CellMask (green). Scale bar, 20 μm . Nuclear intensity of PDX1 and YAP1 are negatively correlated; $r = -0.6$ (**** $P \leq 0.0001$). Each data point corresponds to an individual cell. PDX1/YAP1, $n = 460$; data aggregated from at least three independent experiments.

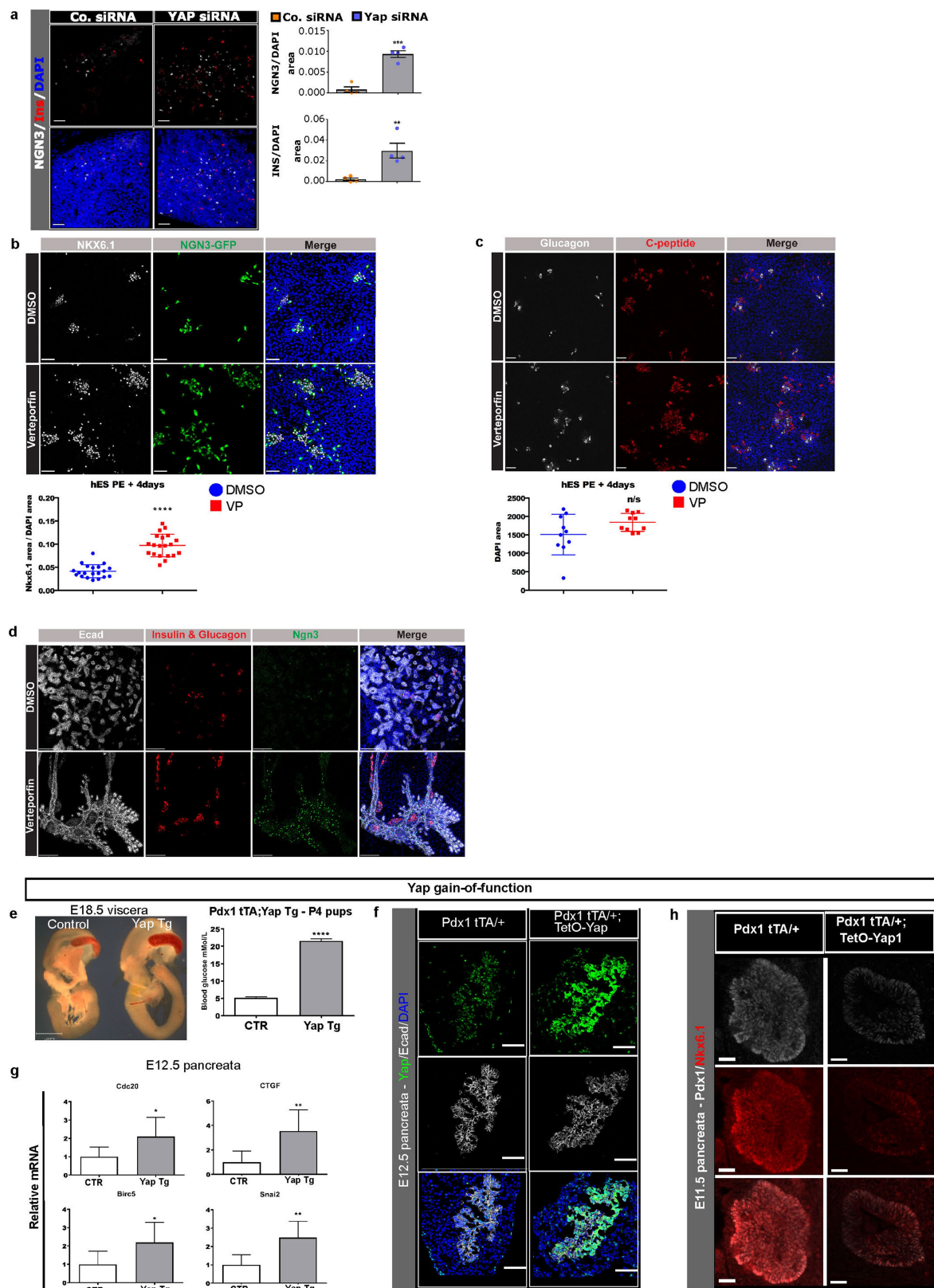


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Loss of YAP1 in pancreatic progenitors

promotes endocrinogenesis in vivo. **a**, qRT-PCR analysis of *Yap1* mRNA expression during developmental and early postnatal stages of pancreas organogenesis. Data represented are compared to E11.5 and normalized to *HPRT* expression. Ordinary one-way ANOVA from three independent embryonic pancreata for each developmental stage; mean \pm s.e.m. **b**, Top, immunofluorescence for YAP1 (red), SOX9 (green) and glucagon (blue) reveals YAP1 expression in SOX9⁺ pancreatic progenitors but not in glucagon⁺ endocrine cells at E10.5, supporting western blot analysis in **c**. Scale bar, 25 μ m. Bottom, co-immunofluorescence analysis for SOX9 (white), NGN3 (green), YAP1 (red) and E-cadherin (E-Cad, blue) on sections of E15.5 pancreas. Nuclear YAP1 expression is seen in SOX9⁺ cells. By contrast, NGN3⁺ endocrine progenitors (arrows) are either YAP1⁻ or YAP1^{low} with nuclear expression. Scale bar, 5 μ m. Images are representative of three embryonic pancreata analysed. **c**, Western blot of E11.5 pancreatic epithelium (epi.) microdissected from mesenchyme (mes.) and analysed for YAP1 protein expression. Mouse adult islets represent mature endocrine tissue. PDX1 served as control for epithelium and islet expression. YAP1 band intensities were normalized to vinculin. Ordinary one-way ANOVA from three independent experiments; mean \pm s.e.m. **d**, qRT-PCR analysis of TEAD isoforms 1–4 and insulin transcripts from mouse embryonic and early postnatal pancreata. Data represented are compared to E11.5 values and normalized to *HPRT* expression. Ordinary one-way ANOVA from three independent embryonic pancreata for each developmental stage; $*P \leq 0.1$; mean \pm s.e.m. **e**, qRT-PCR and western blot analysis of differentiated, sorted hESCs expressing Ngn3–GFP. Data analysed by ordinary one-way ANOVA; data are mean \pm s.e.m. for 3 independent experiments. **f**, Images of dissected gastrointestinal tracts from P4 control (right), and *Pdx1-cre;Yap1^{fl/fl}* (left) littermates. Scale bar, 200 μ m. Image representative of 10 pups per genotype analysed. Pancreas weight of control and *Yap1*

KO pups at P4. $n = 3$ pups analysed. **g**, Scheme depicting the culture of pancreatic explants ex vivo and representative explants whole-mounted and immunostained for E-cadherin (Ecad, grey) and insulin (red). Images represent 3D reconstructions of confocal images using IMARIS. The ratio of insulin⁺ area to total epithelial (Ecad⁺) area was quantified using IMARIS. Explants analysed: wild type, $n = 4$; *Yap1* KO, $n = 7$; $**P = 0.0038$. Scale bar, 30 μ m. qRT-PCR analysis for *Ngn3*. Explants analysed: wild type, $n = 5$; *Yap1* KO, $n = 5$; $*P = 0.0364$ by two-tailed unpaired *t*-test. **h**, Relative gene expression analysis of P4 pancreata for glucagon (*Gcg*) (wild type, $n = 5$; *Yap1* KO, $n = 6$; $*P = 0.0434$), somatostatin (*Sst*) (wild type, $n = 5$; *Yap1* KO, $n = 6$; $**P = 0.0033$), pancreatic polypeptide (*Ppy*) (wild type, $n = 9$; *Yap1* KO, $n = 11$; $***P = 0.0008$) and ghrelin (*Ghrl*) (wild type, $n = 6$; *Yap1* KO, $n = 7$; $***P = 0.0005$) by qRT-PCR. Two-tailed unpaired *t*-test; mean \pm s.d. **i**, qRT-PCR analysis of *Sox9-creER^{T2};Yap1^{fl/fl}* explants from **g** for *Yap1* (wild type, $n = 5$; *Yap1* KO, $n = 11$; $****P = 0.0001$), the endocrine genes insulin (wild type, $n = 8$; *Yap1* KO, $n = 9$; $*P = 0.0100$), glucagon (wild type, $n = 7$; *Yap1* KO, $n = 9$; $***P = 0.004$), somatostatin (wild type, $n = 9$; *Yap1* KO, $n = 8$; $*P = 0.0023$), pancreatic polypeptide (wild type, $n = 8$; *Yap1* KO, $n = 9$; $**P = 0.0013$) and *Ghrl* (wild type, $n = 7$; *Yap1* KO, $n = 7$; $*P = 0.0361$) and pancreatic progenitor genes *Foxa2* (wild type, $n = 7$; *Yap1* KO, $n = 9$; $**P = 0.0030$), *Sox9* (wild type, $n = 7$; *Yap1* KO, $n = 9$; $*P = 0.0206$), *Ptf1a* (wild type, $n = 8$; *Yap1* KO, $n = 8$; $**P = 0.0013$) and *Taz* (wild type, $n = 9$; *Yap1* KO, $n = 11$). Two-tailed unpaired *t*-test; data are mean \pm s.d. **j**, Wild-type E11.5 pancreata cultured on fibronectin-coated dishes for 5 days, treated during the middle 2 days with 1 μ g ml⁻¹ verteporfin (see experimental scheme in **g**) and analysed for endocrine gene expression. Expression represented relative to *HPRT* expression levels compared to DMSO-treated explants. $n = 4$ explants; $*P \leq 0.05$; two-tailed unpaired *t*-test; mean \pm s.d.

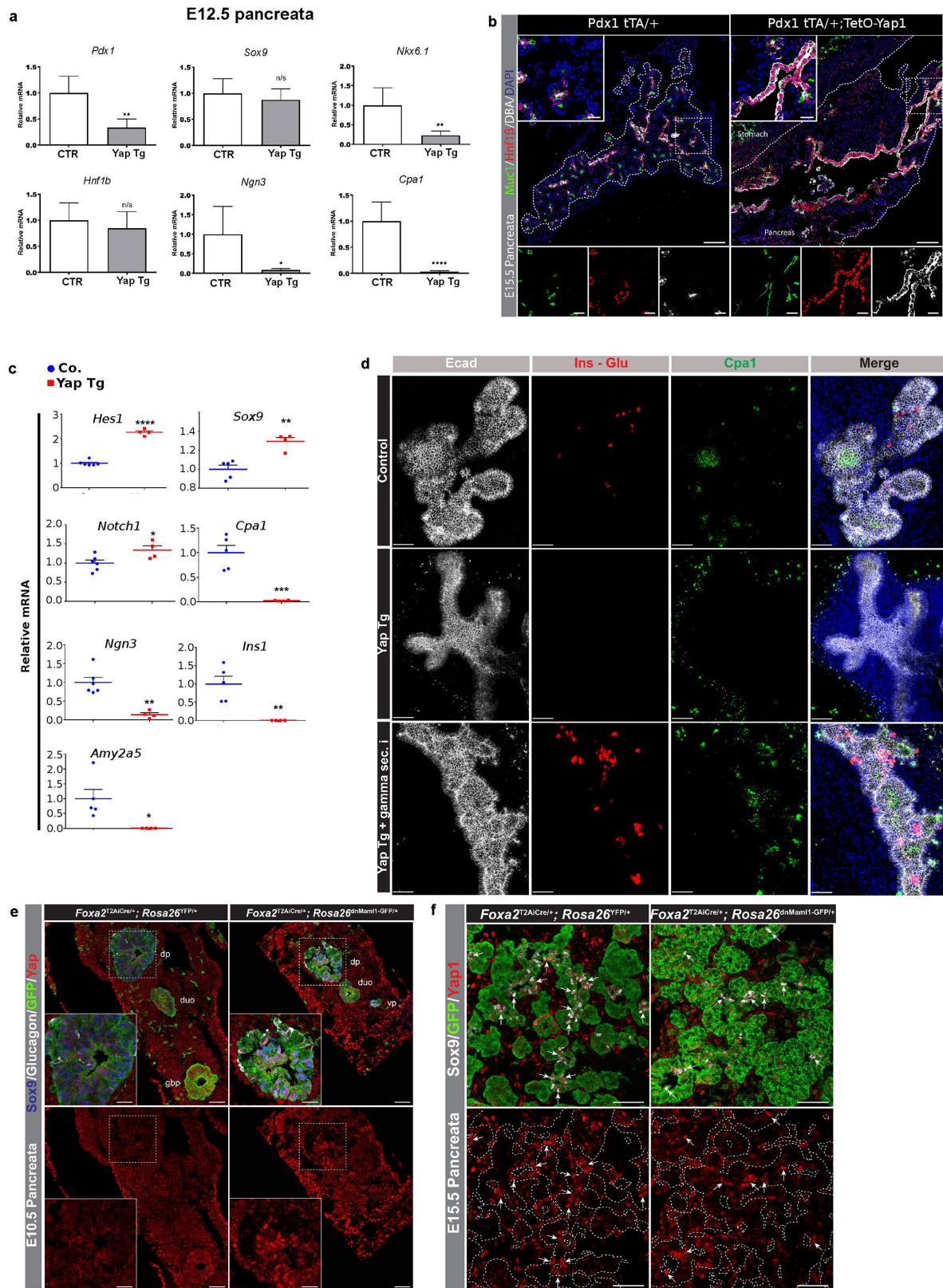


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | YAP1 expression levels are critical in governing mouse and human pancreatic progenitor maintenance.

a, Human in vitro pancreatic progenitor cells were transfected with control (Co) or *YAP1* siRNA on day 13 and fixed after 72 h, then immunostained for NGN3 (grey), insulin (Ins; red) and DAPI (blue). Scale bar, 42 μm . Images and quantifications of NGN3⁺ area/DAPI⁺ area and insulin⁺ area/DAPI⁺ area represent four independent experiments. Two-tailed unpaired *t*-test; mean \pm s.e.m. **b**, hESCs expressing NGN3–GFP, differentiated until S4 stage using the Kieffer protocol and treated with either 1 $\mu\text{g ml}^{-1}$ verteporfin or DMSO for 2 days and left untreated for another 2 days. Immunostaining for NKX6.1 (grey) and GFP (for NGN3) (green), and stained with DAPI (blue). Scale bar, 70 μm . Images are representative of three independent experiments. NKX6.1⁺ area/DAPI⁺ area was quantified using IMARIS. Pooled data from three independent experiments; two-tailed unpaired *t*-test; mean \pm s.d. **c**, Cells prepared as in **b**, immunostained for glucagon (grey) and C-peptide (red) with DAPI (blue). Scale bar, 70 μm . Images are representative of three independent experiments. DAPI⁺ area was quantified using IMARIS. Not significant by two-tailed unpaired *t*-test; mean \pm s.d. **d**, Wild-type E11.5 mouse

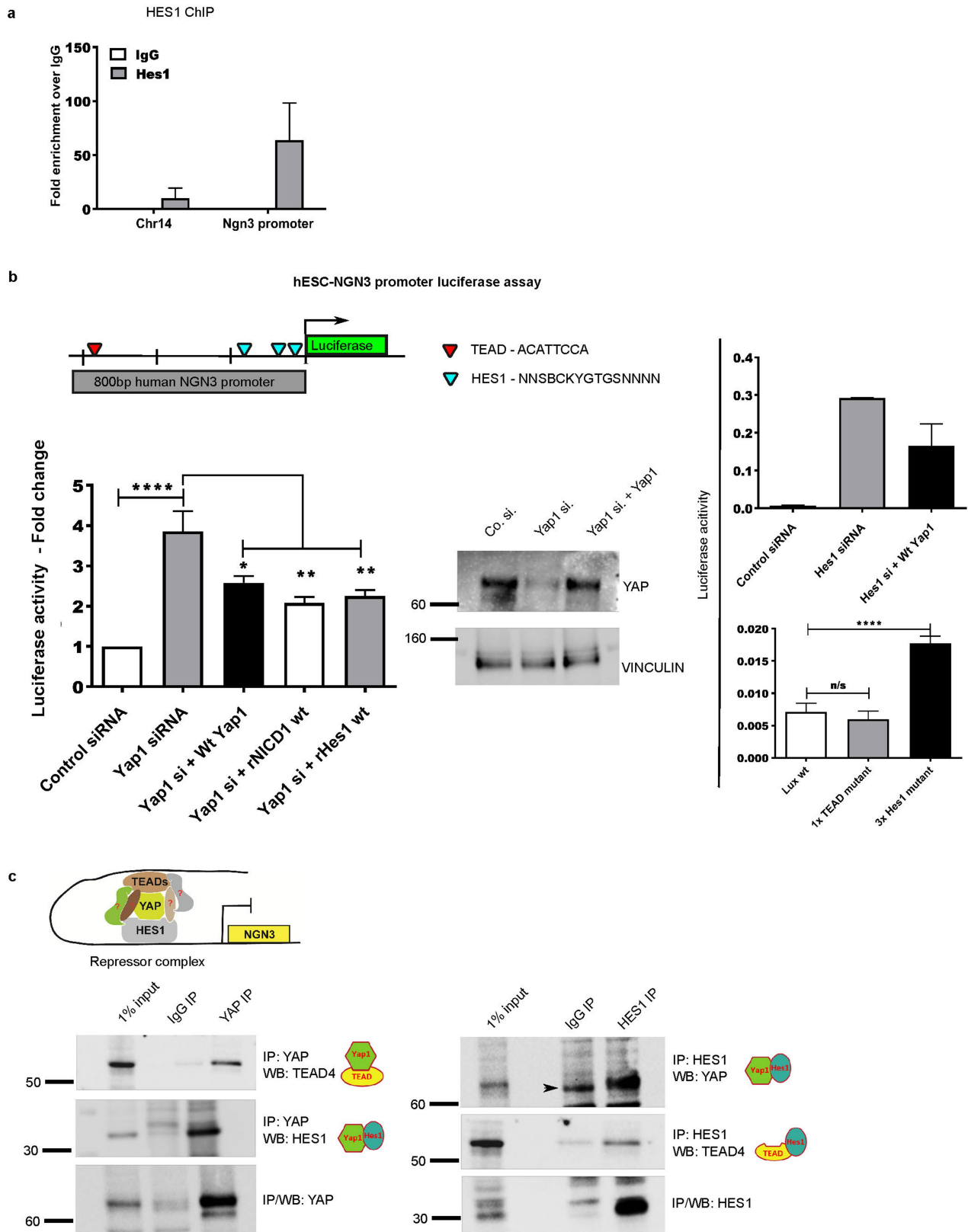
pancreatic explants cultured for 5 days ex vivo with 1 $\mu\text{g ml}^{-1}$ verteporfin or DMSO treatment over days 1–3 and analysed for gene expression after 5 days (Extended Data Fig. 3j). Confocal images of representative explants treated with DMSO ($n = 5$) or verteporfin ($n = 5$), immunostained for E-cadherin (Ecad; grey), insulin and glucagon (red), and NGN3 (green) with DAPI (blue). Scale bar, 100 μm . **e**, E18.5 viscera of control and *tet-YAP1^{S127A};Pdx1^{tTA/+}* (YAP1tg) embryos showing a severe pancreatic agenesis phenotype. YAP1tg P4 pups exhibit severe hyperglycaemia. $n = 3$; two-tailed unpaired *t*-test; mean \pm s.d. **f**, Confocal images of E12.5 pancreata from control and YAP1tg littermates immunostained for E-cadherin (Ecad; grey) and YAP1 (green) with DAPI (blue). Scale bar, 33 μm . Images are representative of three pancreata for each genotype. **g**, qRT–PCR analysis of E12.5 pancreata for known downstream targets of YAP1. Expression is shown relative to HPRT expression. Mean \pm s.d.; $n = 6$ embryos each; two-tailed unpaired *t*-test (* $P \leq 0.05$, ** $P \leq 0.01$). **h**, Whole-mount immunostaining for PDX1 (grey) and NKX6.1 (red) of E11.5 control and YAP1tg pancreata. Scale bar, 30 μm . Images are representative of three control and YAP1tg pancreata.



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Enforced YAP1 expression promotes acquisition of ductal fate upstream of Notch signalling. **a**, qRT-PCR for the pancreatic progenitor markers *Pdx1*, *Sox9*, *Nkx6-1*, *Hnf1b*, *Ngng3* and *Cpa1* in E12.5 control and YAP1tg pancreata. *Pdx1*, *Nkx6-1*, *Ngng3* and *Cpa1* are significantly downregulated, but expression of the ductal progenitors *Sox9* and *Hnf1b* is unchanged. Expression is displayed relative to *Hprt* expression. Mean \pm s.e.m.; $n = 6$ embryos each for control and YAP1tg; two-tailed unpaired *t*-test. **b**, Co-immunofluorescence analysis for Muc1 (green), Hnf1 β (red) and DBA (grey) with DAPI (blue) on sections of E15.5 pancreata from control or YAP1tg littermates. Insets reveal the boxed regions at higher magnification with the channels separated for clarity below. Ducts, as marked by HNF1 β , MUC1 and DBA, are expanded in the YAP1tg (right) compared with control pancreas (left). Scale bar, 100 μ m (main panels) or 25 μ m (insets). **c**, qRT-PCR analysis of E15.5 control (Co.) and YAP1tg pancreata reveals upregulation of *Sox9* (control, $n = 5$; YAP1tg, $n = 4$), *Hes1* (control, $n = 6$; YAP1tg, $n = 4$) and *Notch1* (control, $n = 6$; YAP1tg, $n = 4$) transcripts and significant downregulation of endocrine markers *Ngng3* (control, $n = 6$; YAP1tg, $n = 4$) and *Ins1* (control, $n = 5$; YAP1tg, $n = 4$) and exocrine genes *Cpa1* (control, $n = 6$; YAP1tg, $n = 6$) and *Amy2a5* (control, $n = 5$; YAP1tg, $n = 4$). * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.0001$; two-tailed unpaired *t*-test; mean \pm s.e.m. **d**, Whole-mount immunostaining of E11.5 + 4 day control

and YAP1tg explants cultured and treated with DMSO (control and YAP1tg) or 25 μ M γ -secretase inhibitor (YAP1tg). Immunostaining with antibodies against E-cadherin (Ecad; white), insulin (ins) and glucagon (glu) (red) to indicate endocrine differentiation, and CPA1 (green) for acinar differentiation, with DAPI (blue). Scale bar, 60 μ m. Note that there is non-specific background with the CPA1 (green) staining outside the epithelium (especially in YAP1tg). Images are representative of $n = 5$ explants analysed for each condition. **e**, Confocal optical sections of E10.5 pancreata from *Foxa2*^{T2AiCre/+};*Rosa26*^{YFP/+} control and Notch signalling mutant (*Foxa2*^{T2AiCre/+};*Rosa26*^{dnMaml1-eGFP/+}) embryos stained for SOX9 (blue), glucagon (grey), GFP (green: shows recombination efficiency in epithelium) and YAP1 (red). Scale bar, 50 μ m (main panels), 25 μ m (insets). dnMaml1-induced Notch blockade in the pancreatic endoderm did not affect YAP1 expression in SOX9⁺ progenitors. Images are representative of three pancreata analysed from each genotype. **f**, YAP1 is normally expressed in the epithelium (GFP⁺, green) and mesenchyme of E15.5 secondary transition control (*Foxa2*^{T2AiCre/+};*Rosa26*^{YFP/+}) pancreas and its expression in SOX9⁺ progenitors is unaffected in Notch signalling mutants (*Foxa2*^{T2AiCre/+};*Rosa26*^{dnMaml1-eGFP/+}). SOX9 (grey), YAP1 (red) and GFP (green, shows recombination efficiency in epithelium). Scale bar, 50 μ m. Images are representative of three pancreata analysed from each genotype.

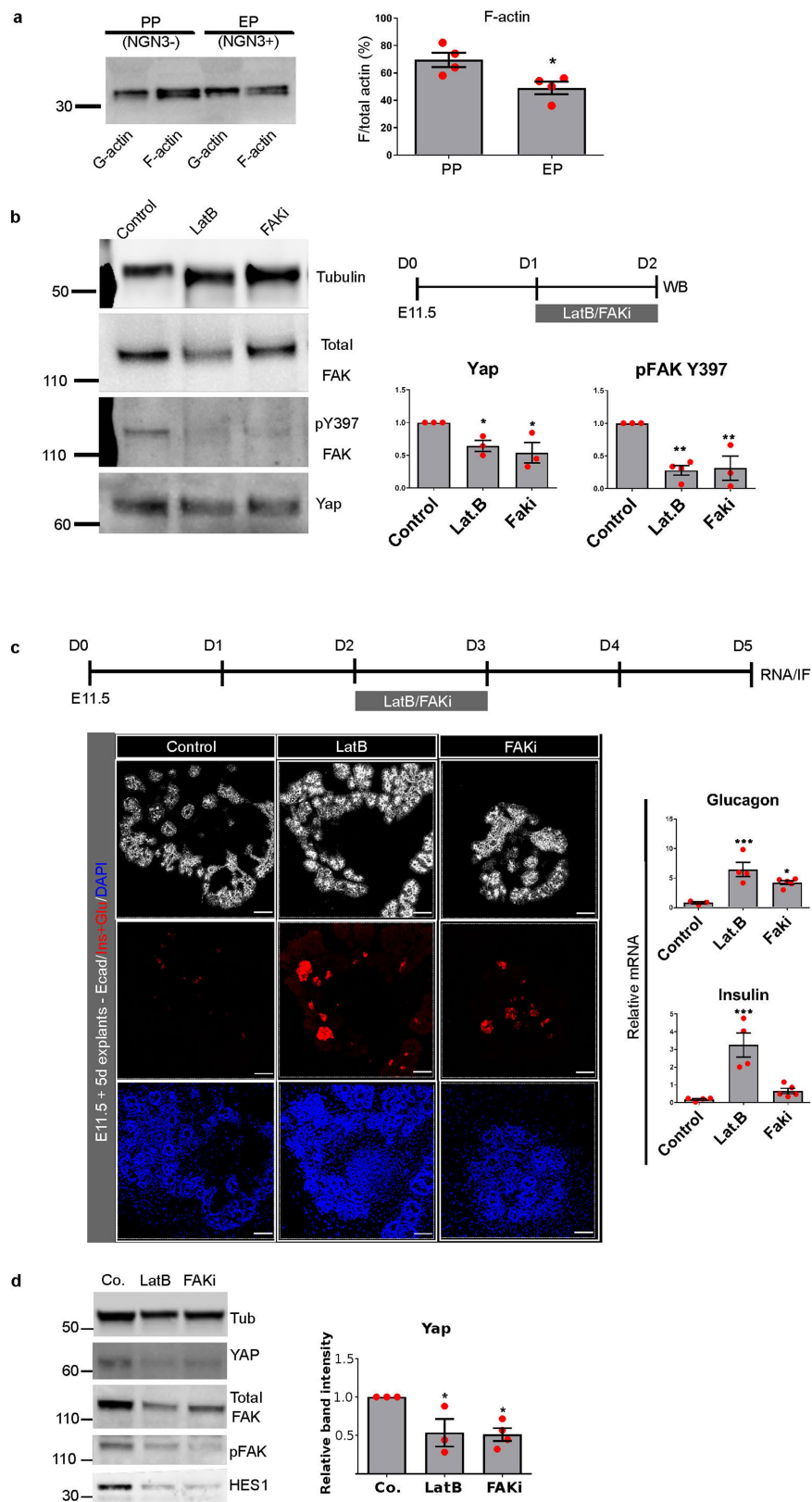


Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | YAP1 is an essential mediator of Notch-dependent *NGN3* repression **a**, ChIP-qPCR of a site 250 bp upstream of the ATG start site of the human *NGN3* locus using HES1 antibody. HES1 is specifically enriched at the *NGN3* promoter in human in vitro pancreatic progenitor cells, in which *NGN3* is repressed. Data represent fold-enrichment over IgG pull-down. Chr14 is used as genomic negative control. Data are mean \pm s.e.m. of three independent experiments. **b**, *NGN3* promoter luciferase assay. Left, knockdown of endogenous YAP1 in in vitro pancreatic progenitor cells upregulates *NGN3* promoter-driven luciferase activity and the overexpression of exogenous wild-type YAP1, wild-type rat NICD1 or wild-type rat HES1 partially rescues the effects of YAP1 knockdown. $n = 3$ experiments; $*P \leq 0.05$, $**P \leq 0.01$, $****P \leq 0.0001$; ordinary one-way ANOVA; data are mean \pm s.e.m. Middle, western blot analysis of siRNA-treated samples for YAP1 and vinculin loading control. Top right, loss of HES1 results in upregulation

of luciferase activity and is partially rescued by overexpression of wild-type YAP1. $n = 2$ experiments; data are mean \pm s.e.m. Bottom right, point mutation of TEAD binding site does not affect luciferase activity, indicating that the binding of YAP1-TEAD repressor complex is not mediated through this site, but the $3\times$ HES1 binding-site mutation significantly upregulates promoter activity. Ordinary one-way ANOVA; n.s, not significant; data are mean \pm s.e.m.; $n = 6$ experiments.

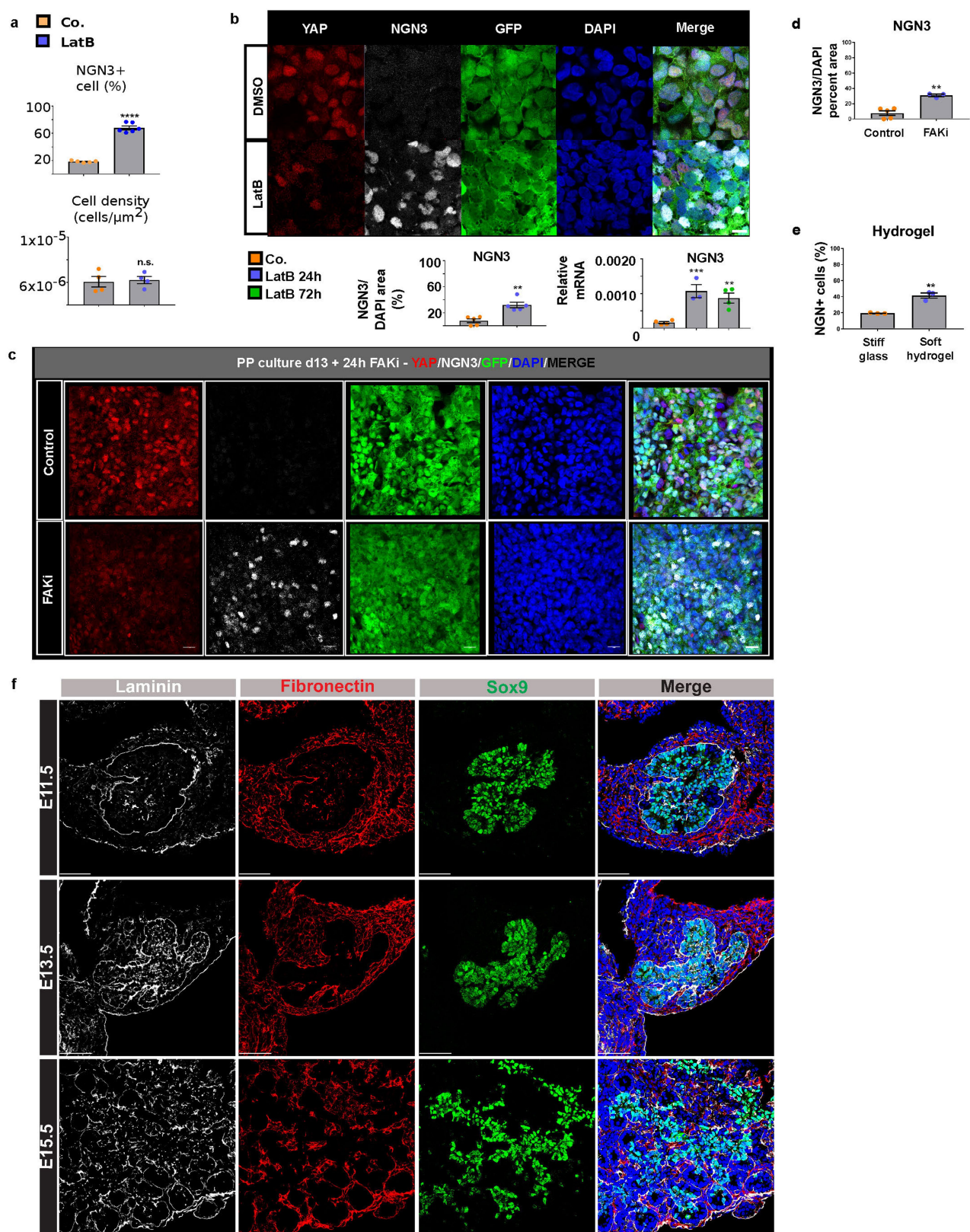
c, Co-immunoprecipitation in in vitro pancreatic progenitors. YAP1 forms an endogenous protein complex with TEAD4 and HES1. Likewise, HES1 forms a protein complex with TEAD4 and YAP1, supporting ChIP data on the *NGN3* promoter in the pancreatic progenitors. Arrowhead indicates non-specific IgG band. Cartoon depicting a possible YAP1-TEAD-HES1-based transcriptional repressor complex at the *NGN3* locus. Representative blots from three independent co-immunoprecipitations performed.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Actin and FAK signalling modulate pancreatic progenitor cell fate via YAP1. **a**, Representative western blots for F-actin and G-actin from hESC-derived NGN3⁺ (representing endocrine precursor cells) and NGN3⁻ (representing pancreatic progenitor cells) sorted at the endocrine precursor stage. * $P \leq 0.05$; two-tailed unpaired t -test; data are mean \pm s.e.m.; $n = 4$ independent experiments. **b**, Scheme depicting timeline of pancreas explant culture and treatments. Immunoblots and quantifications of tubulin, FAK, pFAK Thr397 and YAP1 proteins in explants treated with DMSO, latB or FAK inhibitor after 24 h. YAP1 and pFAK band intensities are normalized to tubulin. Representative images of 3 independent western blot analyses. * $P \leq 0.05$, ** $P \leq 0.01$; two-tailed unpaired t -test; data are mean \pm s.e.m. **c**, Schematic of explant culture timeline. Single-plane optical section confocal images

of whole-mount immunostaining for E-cadherin (Ecad, grey) and insulin and glucagon (red) with DAPI (blue). Scale bar, 100 μ m. Images are representative of 5 explants analysed for each condition. Gene expression analysis from similar experiments for glucagon and insulin expression. Expression is shown relative to *HPRT* expression; * $P \leq 0.05$, *** $P \leq 0.001$; ordinary one-way ANOVA; mean \pm s.e.m. Insulin control, $n = 4$; latB, $n = 4$; FAK inhibitor, $n = 5$; glucagon control, $n = 3$; latB, $n = 4$; FAK inhibitor, $n = 5$. **d**, Western blot analysis of cells represented in (Extended Data Fig. 8b). Cells treated with latB (1 μ M) or FAK inhibitor (3 μ M) for 24 h. YAP1 band intensity is normalized to tubulin of control cells. * $P \leq 0.05$; two-tailed unpaired t -test; data are mean \pm s.e.m.; $n = 3$ independent experiments.

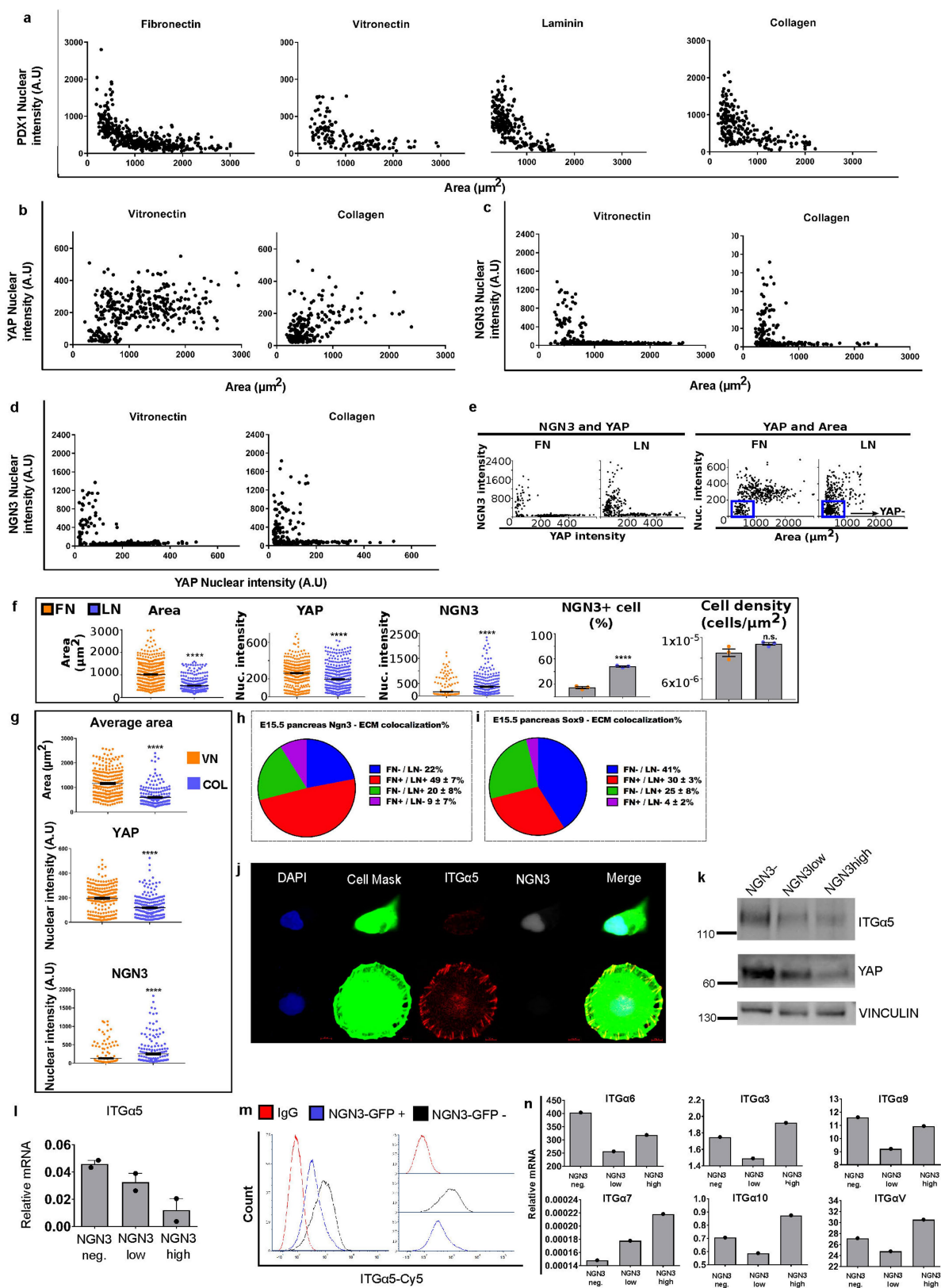


Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Endocrine differentiation is negatively influenced by FAK and actin signalling, and is associated with increased laminin expression in vivo. **a**, Pancreatic progenitors sorted on a dish at single-cell density treated (as in Fig. 4c) with the F-actin inhibitor latB (1 μ M) or control (DMSO). Single cells stained for YAP1 and NGN3 and quantified as in Fig. 1a. Each data point representing NGN3⁺ cells and density corresponds to one experiment. Percentage of NGN3⁺ cell number represents the number of NGN3⁺ cells divided by the total number of DAPI⁺ cells. n.s., non-significant. Two-tailed unpaired *t*-test; ****P* \leq 0.001; data are mean \pm s.e.m. **b**, Unsorted pancreatic progenitor culture treated with DMSO (control) or latB for 24 h and stained for YAP1 (red), NGN3 (grey) and GFP-PDX1 (green) with DAPI (blue). Scale bar, 20 μ m. Representative images of three independent experiments. Image quantification of NGN3 protein is a measure of NGN3⁺/DAPI⁺ per cent area. Each data point representing NGN3⁺ cells and density corresponds to one experiment. ***P* \leq 0.01; ****P* \leq 0.001; two-tailed unpaired *t*-test; data are mean \pm s.e.m. Gene expression of NGN3. Each data point corresponds to one experiment. Ordinary one-way ANOVA;

mean \pm s.e.m. **c**, Unsorted pancreatic progenitor cultures treated with DMSO (control) or FAK inhibitor for 24 h then stained for YAP1 (red), NGN3 (grey), GFP-PDX1 (green) with DAPI (blue). Scale bar, 20 μ m. Images are representative of three independent experiments.

d, Quantification of **c**. NGN3 protein quantification following FAK inhibitor treatment calculated as NGN3⁺ area/DAPI⁺ per cent area. Two-tailed unpaired *t*-test; ***P* \leq 0.01; mean \pm s.e.m. Each data point corresponds to one experiment. **e**, Sorted single pancreatic progenitor cells plated on soft substrate (10 kPa) versus glass (\geq 2,000,000 kPa) coated with fibronectin for 24 h and analysed for percentage NGN3 expression as calculated in **a**. Percentage of NGN3⁺ cell number is calculated as the number of NGN3⁺ cells divided by total number of DAPI⁺ cells. Data are mean \pm s.e.m. of *n* = 3 experiments; two-tailed unpaired *t*-test; ***P* \leq 0.01. **f**, Co-immunofluorescence analysis of 30- μ m sections from E11.5, E13.5 and E15.5 pancreata for laminin (grey), fibronectin (red) and SOX9 (green) with DAPI (blue). Scale bar, 53 μ m. Images are representative of three wild-type pancreata analysed.

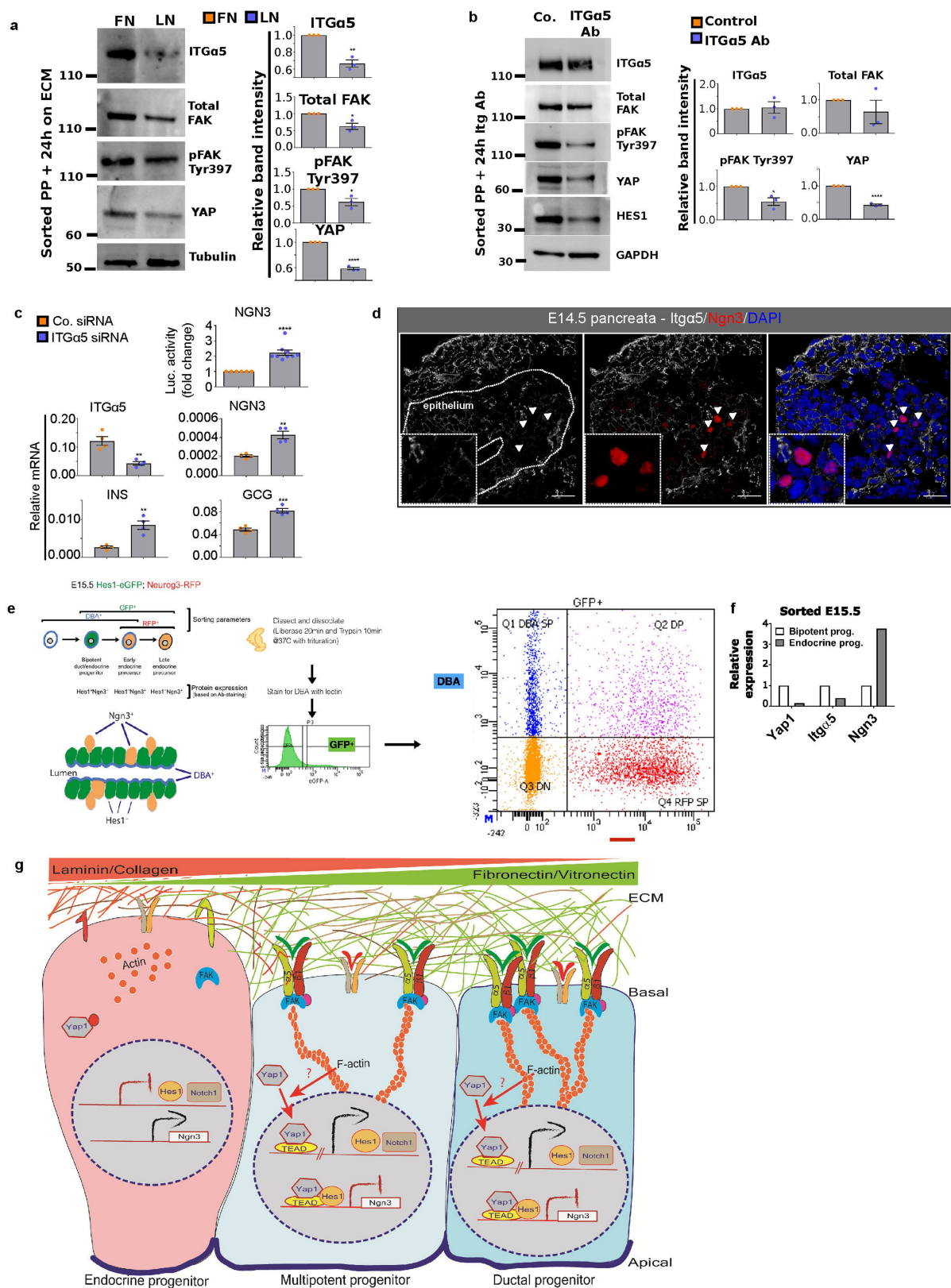


Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Endocrine differentiation is modulated by ECM composition-mediated integrin signalling.

a–e, Pancreatic progenitors sorted at single-cell density on dishes coated with fibronectin, laminin, vitronectin or collagen, fixed after 24 h and stained for PDX1 (fibronectin, $n = 436$; laminin, $n = 252$; vitronectin, $n = 133$; collagen, $n = 227$ cells), YAP1 (vitronectin, $n = 430$; collagen, $n = 316$ cells), NGN3 (vitronectin, $n = 300$; collagen, $n = 205$ cells) or YAP1–NGN3 nuclear intensity (vitronectin, $n = 300$; collagen, $n = 216$ cells) were quantified as in Fig. 1a. Small blue boxes in **e** show the YAP1[−] cell population quantified. Graphs present plots of individual cell data aggregated from three independent experiments. **f**, Comparison quantification, as shown in Fig. 4c. Each data point representing YAP1⁺, NGN3⁺ cells and density corresponds to mean \pm s.e.m. of three independent experiments. Percentage of NGN3⁺ cell number was determined as in Extended Data Fig. 8a. Two-tailed unpaired t -test; **** $P \leq 0.0001$, n.s., non-significant; data are mean \pm s.e.m. **g**, Pancreatic progenitor cells sorted as in Fig. 4c and analysed after 24 h culture on vitronectin or collagen. Cell area and nuclear intensities of YAP1⁺ and NGN3⁺ cells were measured for cells plated on both ECMs. Two-tailed unpaired t -test; **** $P \leq 0.0001$, mean \pm s.e.m. of cell data aggregated from three independent experiments. **h, i**, Co-immunofluorescence and quantification from 30- μ m sections from E15.5 wild-type pancreata co-stained for NGN3, fibronectin and laminin (**h**). NGN3⁺ cells were segmented using IMARIS surface module. Identification of cells in contact with either laminin only or fibronectin only or both was computed using percentage intensity overlap. The threshold for the intensity overlap was selected manually by examining individual cells. Pie chart showing the percentage of segmented NGN3⁺ cells in contact with fibronectin or laminin or both. Shown is the mean percentage and confidence interval data of 1,390 NGN3⁺ cells from 4 independent wild-type E15.5 embryos analysed.

i, Pie chart showing the percentage of segmented Sox9⁺ cells in contact with fibronectin or laminin or both. Shown is the mean percentage and confidence interval data of 2,637 Sox9⁺ cells from 4 independent wild-type E15.5 embryos analysed. **j**, Sorted single pancreatic progenitor cells confined (top) or spread (bottom), stained for integrin $\alpha 5$ (ITGa5; red) and NGN3 (grey), with CellMask (green) and DAPI (blue). Scale bar, 10 μ m. Representative images from three independent experiments are shown. **k**, Western blots from sorted hESC-derived cells expressing NGN3–GFP as in Extended Data Fig. 3e. Pancreatic progenitor differentiation performed with the NGN3–GFP reporter line. NGN3^{high} and NGN3^{low} cells (endocrine precursors) express significantly lower levels of integrin $\alpha 5$ (ITGa5) and YAP1 protein than NGN3[−] cells (pancreatic progenitors). Representative blots from two independent experiments shown. **l**, Integrin $\alpha 5$ (*ITGA5*) mRNA expression analysis of NGN3[−], NGN3^{low} and NGN3^{high} populations derived from Extended Data Fig. 3e. *ITGA5* expression mirrors *YAP1* expression during endocrine differentiation (Extended Data Fig. 3e). Expression is presented relative to *GAPDH* expression. Data are mean \pm s.e.m. from two independent differentiation experiments. **m**, IgG or integrin $\alpha 5$ (ITGa5) antibody staining and flow cytometric analysis of differentiated hESC NGN3–GFP reporter cells. Left, histogram overlay and mean fluorescence intensities (MFI) of IgG-stained (MFI = 100), NGN3–GFP⁺ endocrine precursor (MFI = 310) and NGN3–GFP[−] pancreatic progenitor cell (MFI = 704) fractions. Right, staining distribution and MFI of each cell population. NGN3⁺ endocrine precursors display lower protein expression levels of integrin $\alpha 5$ compared to NGN3[−] pancreatic progenitors. Data represent three independent experiments. **n**, mRNA expression analysis of other α integrins (as in **l**) that remain unchanged during differentiation of NGN3⁺ cells from pancreatic progenitors. Expression is shown relative to *GAPDH* expression; $n = 1$ experiment.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Pancreatic progenitor cell fate specification is regulated via an ECM–integrin–FAK–actin–YAP1–Notch signalling cascade. **a**, Western blot of sorted pancreatic progenitor cells after 24 h adhesion on fibronectin (FN) or laminin (LN). Band intensities were normalized against tubulin. $n = 3$ independent experiments. pFAK and FAK protein levels are reduced by a similar extent on laminin, indicating that the fraction of pFAK remains constant. Two-tailed unpaired t -tests; $*P \leq 0.05$; $**P \leq 0.01$, $***P \leq 0.001$; mean \pm s.e.m. **b**, Western blot of sorted pancreatic progenitor cells after 24 h adhesion with control or integrin $\alpha 5$ (ITGa5) inhibition. Band intensities were normalized against GAPDH intensity of control antibody-treated cells. $n = 3$ independent experiments. pFAK and FAK protein levels are reduced by a similar extent by laminin treatment, indicating that the fraction of pFAK remains constant. Two-tailed unpaired t -tests; $*P \leq 0.05$, $***P \leq 0.001$; mean \pm s.e.m. **c**, Unsorted pancreatic progenitor culture transiently transfected with integrin $\alpha 5$ (ITGA5) or control siRNA and analysed after 72 h. Top, NGN3 promoter luciferase assay. Right, gene expression of ITGA5, NGN3, INS and GCG. Each data point corresponds to one

experiment. Two-tailed t -tests, $**P \leq 0.01$, $***P \leq 0.001$, $****P \leq 0.0001$; mean \pm s.e.m. **d**, Expression of integrin $\alpha 5$ in vivo during secondary transition of mouse embryonic pancreas development. Wild-type E14.5 pancreas stained for integrin $\alpha 5$ (ITGa5; grey) and NGN3 (red) with DAPI (blue). Scale bar, 20 μ m. Images are representative of three wild-type pancreata analysed. **e**, Schematic of sorting mouse E15.5 bi-PPs and endocrine cell populations: *Hes1-eGFP*; *Nggn3-tRFP* double reporter mice were dissected and dissociated, stained with DBA and sorted by FACS. Bi-PPs were represented by the $\text{DBA}^+ \text{NGN3-tRFP}^- \text{HES1-EGFP}^+$ population and endocrine precursors were represented by the $\text{DBA}^+ \text{NGN3-tRFP}^+ \text{HES1-EGFP}^+$ population. All NGN3-tRFP^+ cells were also HES1-EGFP^+ owing to the long half-life of EGFP. **f**, Gene expression from microarray analysis of E15.5 FACS-sorted bi-PPs, as in Extended Data Fig. 10e. White bar, $\text{DBA}^+ \text{NGN3-tRFP}^- \text{HES1-EGFP}^+$; grey bar, $\text{DBA}^+ \text{NGN3-tRFP}^+ \text{HES1-EGFP}^+$. Shown are all significant differentially expressed (adjusted P values < 0.01); $n = 4$; pools of 5 pancreata. **g**, Model for bi-PP cell fate specification regulated via ECM–integrin–FAK–actin–YAP1–Notch signalling cascade.

VCAM-1⁺ macrophages guide the homing of HSPCs to a vascular niche

Dantong Li^{1,2,10}, Wenzhi Xue^{1,10}, Mei Li^{1,2,10}, Mei Dong¹, Jianwei Wang¹, Xianda Wang¹, Xiyue Li¹, Kai Chen¹, Wenjuan Zhang¹, Shuang Wu¹, Yingqi Zhang³, Lei Gao^{1,4}, Yujie Chen⁵, Jianfeng Chen⁶, Bo O. Zhou⁶, Yi Zhou⁷, Xuebiao Yao⁸, Lin Li⁶, Dianqing Wu⁹ & Weijun Pan^{1,2*}

Haematopoietic stem and progenitor cells (HSPCs) give rise to all blood lineages that support the entire lifespan of vertebrates¹. After HSPCs emerge from endothelial cells within the developing dorsal aorta, homing allows the nascent cells to anchor in their niches for further expansion and differentiation^{2–5}. Unique niche microenvironments, composed of various blood vessels as units of microcirculation and other niche components such as stromal cells, regulate this process^{6–9}. However, the detailed architecture of the microenvironment and the mechanism for the regulation of HSPC homing remain unclear. Here, using advanced live imaging and a cell-labelling system, we perform high-resolution analyses of the HSPC homing in caudal haematopoietic tissue of zebrafish (equivalent to the fetal liver in mammals), and reveal the role of the vascular architecture in the regulation of HSPC retention. We identify a VCAM-1⁺ macrophage-like niche cell population that patrols the inner surface of the venous plexus, interacts with HSPCs in an ITGA4-dependent manner, and directs HSPC retention. These cells, named ‘usher cells’, together with caudal venous capillaries and plexus, define retention hotspots within the homing microenvironment. Thus, the study provides insights into the mechanism of HSPC homing and reveals the essential role of a VCAM-1⁺ macrophage population with patrolling behaviour in HSPC retention.

In vertebrates, the establishment of the HSPC pool is a dynamic process that requires not only the HSPC fate specification from the haemogenic endothelium, but also their subsequent homing to distinct anatomic sites^{2–5}. In the zebrafish, HSPCs are initially formed in the ventral wall of the dorsal aorta in the aorta–gonad–mesonephros (AGM) region^{3,4}. The nascent HSPCs then migrate to the caudal haematopoietic tissue (CHT) and kidney marrow, which are the haematopoietic tissues equivalent to mammalian fetal liver and bone marrow, respectively, where the HSPCs undergo rapid expansion and differentiation to support larval and adult haematopoiesis^{2,10}. However, how HSPCs migrate to and finally colonize these tissues remains poorly understood.

To investigate these unknown mechanisms, we carried out a large-scale forward genetics screen in zebrafish for mutants that display HSPC homing defects. The mutant line cas005 showed severe defects in definitive haematopoiesis, but normal primitive haematopoiesis and vascular morphogenesis (Extended Data Fig. 1a, b, e, g). Although the haemogenic endothelium in *mut^{cas005}* was intact, as revealed by whole-mount in situ hybridization (WISH) results of the nascent HSPC marker *runx1*¹¹ (a key transcription factor that regulates nascent HSPC emergence), the number of HSPCs in the mutant CHT was severely

reduced (Extended Data Fig. 1c–e, g), without increased HSPC apoptosis (Extended Data Fig. 1f) compared to the wild-type CHT.

The genetic mutation was mapped to a loss-of-function mutation in the *integrin alpha 4 (itga4)* gene by positional cloning (Extended Data Fig. 2a–c). Indeed, morpholino-mediated knockdown of *itga4* expression (Extended Data Fig. 2e–g) or a second zebrafish *itga4^{cas010}* mutant generated by CRISPR–Cas9¹² (Extended Data Fig. 2b–d) displayed similar phenotypes to that of *mut^{cas005}*, which was hence renamed as *itga4^{cas005}*.

WISH analysis showed that *itga4* expression was enriched in both the AGM and the CHT in a *runx1*¹¹- and *myb*¹³ (another key transcription factor that regulates nascent HSPC migration into circulation)-dependent manner (Extended Data Fig. 2h). Conversely, *runx1* enhancer¹⁴-directed definitive HSPC re-expression of wild-type *itga4* could rescue the *itga4* mutant defects (Extended Data Fig. 2i–k), indicating an HSPC cell-autonomous role of ITGA4.

The defective definitive haematopoiesis in zebrafish *itga4* mutants is consistent with a previous report¹⁵. The VLA-4 integrin, composed of α_4 (*itga4*) and β_1 (*itgb1*) subunit, is predominantly expressed on HSPCs in mammals in early embryogenesis¹⁶. In mice, the α_4 integrin is essential for normal haematopoietic development in the fetal liver^{15,17}, and inhibition of α_4 could mobilize HSPCs from fetal livers by interfering with the homing and retention process^{18,19}. However, the precise mechanism by which ITGA4 regulates HSPC homing remains largely unknown.

To achieve real-time characterization of HSPC homing to, and retention in, the CHT, we took advantage of the transgenic line *Tg(kdrl:Dendra2)*, in which the *kdrl* gene promoter drives the expression of the photoconvertible Dendra2 fluorescent protein in the entire vasculature. At 36 hours post-fertilization (h.p.f.), we converted the green fluorescence of Dendra2⁺ endothelial cells in the AGM to red (Extended Data Fig. 3a). Consistent with previous reports^{3,4}, a substantial number of endothelial cells converted by endothelial-to-haematopoietic transition emerged from the aortic ventral wall into the sub-aortic space, subsequently entered the blood circulation, and finally colonized the CHT by 48–50 h.p.f. (Extended Data Fig. 3a, b).

These photoconverted red Dendra2⁺ cells in the CHT were found to carry *runx1* transcripts (Extended Data Fig. 3b). In addition, the knockdown of *runx1* or *myb* expression¹³ significantly reduced the number of photoconverted red Dendra2⁺ cells in the CHT (Extended Data Fig. 3c, d). These results confirmed that the photoconverted red Dendra2⁺ cells homing to the CHT were nascent HSPCs.

Thus, we were able to characterize the entire process and individual HSPC homing–retention events in the CHT. We found that the

¹Key Laboratory of Tissue Microenvironment and Tumor, CAS Center for Excellence in Molecular Cell Science, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, University of Chinese Academy of Sciences, Chinese Academy of Sciences (CAS), Shanghai, China. ²Key Laboratory of Stem Cell Biology, Shanghai Jiao Tong University School of Medicine (SJTUSM) & Shanghai Institutes for Biological Sciences (SIBS), CAS, Shanghai, China. ³Department of Orthopedic Surgery, Tongji Hospital, Tongji University School of Medicine, Shanghai, China. ⁴Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, CAS, Beijing, China. ⁵Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, University of Chinese Academy of Sciences, CAS, Shanghai, China. ⁶State Key Laboratory of Molecular Biology, CAS Center for Excellence in Molecular Cell Science, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, CAS, Shanghai, China. ⁷Stem Cell Program and Division of Hematology/Oncology, Boston Children's Hospital and Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. ⁸CAS Center for Excellence in Molecular Cell Science, University of Science and Technology of China, Hefei, China. ⁹Department of Pharmacology, Vascular Biology and Therapeutic Program, School of Medicine, Yale University, New Haven, CT, USA. ¹⁰These authors contributed equally: Dantong Li, Wenzhi Xue, Mei Li. *e-mail: weijunpan@sibs.ac.cn

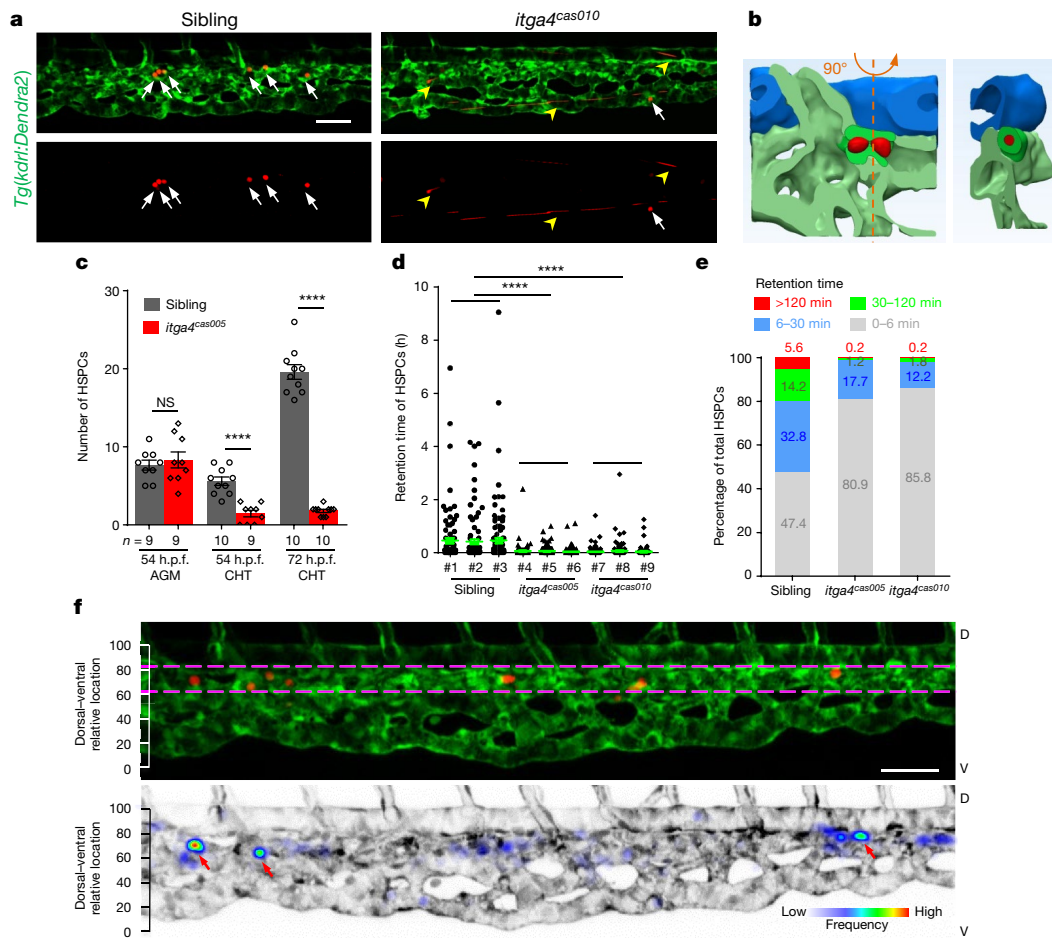


Fig. 1 | Live-imaging characterization of nascent HSPCs retention in the CHT. **a**, Frame shots from the CHT at 54 h.p.f. show HSPCs seeding successfully (white arrows) in wild-type siblings but not in *itga4^{cas010}* mutants (fast moving, yellow arrowheads). See Supplementary Video 1. **b**, A representative vascular architecture view of the HSPC retention hotspot. The orthogonal view is shown on the right. HSPCs, red; dorsal aorta, blue; venous plexus, light green; venous capillary, dark green. See Supplementary Video 4. **c**, The number of HSPCs in the AGM and CHT of *itga4^{cas005}* mutants and wild-type siblings at 54 and 72 h.p.f., respectively. 54 h.p.f. AGM: $P = 0.59$, $t = 0.55$, $df = 16$; 54 h.p.f. CHT: **** $P < 0.0001$, $t = 6.00$, $df = 17$; 72 h.p.f. CHT: **** $P < 0.0001$, $t = 18.65$, $df = 18$. NS, not significant. **d**, **e**, Retention time of individual HSPCs in each embryo (**d**) and percentage of total HSPCs in four classified retention time zones in group embryos ($n = 3$) (**e**) of wild-type siblings and *itga4^{cas005}* and *itga4^{cas010}* mutants during 50–60 h.p.f. Wild type vs *itga4^{cas005}*: $P < 0.0001$, $t = 8.56$, $df = 824$; wild type vs *itga4^{cas010}*: $P < 0.0001$, $t = 7.93$, $df = 758$. **f**, Top, HSPCs that remained for longer than 30 min were preferentially located in the region enriched with venous capillaries (between the two magenta dashed lines). Bottom, the frequency of the appearance of HSPCs in the entire CHT of *Tg(kdrl:Dendra2)* zebrafish larvae from 50 to 60 h.p.f. (retention hotspots marked by red arrows). D, dorsal; V, ventral. Scale bars, 50 μm (**a**, **f**).

lodgement of HSPCs initially took place at approximately 48–50 h.p.f., and the number of lodged HSPCs markedly increased over 24 h. However, in the *itga4*-mutant embryos, HSPC retention was barely detectable (Fig. 1a, c and Supplementary Video 1). More specifically, the average retention time of HSPCs in wild-type embryos is close to 30 min, whereas the HSPCs in the *itga4* mutants went through the CHT quickly with very short retention times (average retention time of 4 min) (Fig. 1d, e, Extended Data Fig. 3e and Supplementary Video 2). The functional consequence of the disrupted HSPC retention event in the CHT as the lodgement of HSPCs for a period of more than 30 min (over 20% of HSPCs in the wild-type embryos, but less than 2% of HSPCs in the *itga4* mutants).

It has been proposed that HSPCs reside in the anatomically defined niche, where they receive and integrate regulatory signals from the niche cells and extracellular matrices^{6,7,20} for their expansion and differentiation. To understand the tendency of HSPC retention, we traced individual HSPCs and correlated their retention time with the dorsal–ventral relative location. This scatterplot analysis revealed that the longer HSPCs resided in the CHT, the greater was their tendency

not significant. **d**, **e**, Retention time of individual HSPCs in each embryo (**d**) and percentage of total HSPCs in four classified retention time zones in group embryos ($n = 3$) (**e**) of wild-type siblings and *itga4^{cas005}* and *itga4^{cas010}* mutants during 50–60 h.p.f. Wild type vs *itga4^{cas005}*: $P < 0.0001$, $t = 8.56$, $df = 824$; wild type vs *itga4^{cas010}*: $P < 0.0001$, $t = 7.93$, $df = 758$. **f**, Top, HSPCs that remained for longer than 30 min were preferentially located in the region enriched with venous capillaries (between the two magenta dashed lines). Bottom, the frequency of the appearance of HSPCs in the entire CHT of *Tg(kdrl:Dendra2)* zebrafish larvae from 50 to 60 h.p.f. (retention hotspots marked by red arrows). D, dorsal; V, ventral. Scale bars, 50 μm (**a**, **f**).

to reach the dorsal part of the caudal venous plexus (CVP) (Extended Data Fig. 4a). Next, we analysed the frequency of HSPCs' appearance in the entire CHT over an 8-h time period. Unexpectedly, the retention of HSPCs was not evenly distributed in the dorsal part of the CVP, and instead the cells frequently occurred in several regions of the CHT in wild-type embryos. We referred to these regions as the retention 'hot-spots' (Fig. 1f and Extended Data Fig. 4b, c). These retention hotspots are largely localized at the venous capillary confluence points that are connected to the CVP, in which the velocity of circulating photoconverted HSPCs is notably reduced (Extended Data Fig. 5a). HSPCs that entered the CHT either from the intersegmental vessel (ISV) or from the CVP (Extended Data Fig. 5b–e and Supplementary Video 3) were sharply decelerated. We also found that most HSPCs that remained for more than 30 min were in the venous capillaries, which have similar diameters to that of HSPCs (Fig. 1b, Extended Data Fig. 6a–f, h, Supplementary Video 4). In the *itga4* mutants, the retention hotspots were not evident (Extended Data Fig. 4d). However, there was no significant difference in the number, size or confluence points of the vascular architecture between wild-type embryos and *itga4* mutants (Extended Data Fig. 6g, h). These observations led us to hypothesize that other niche components might

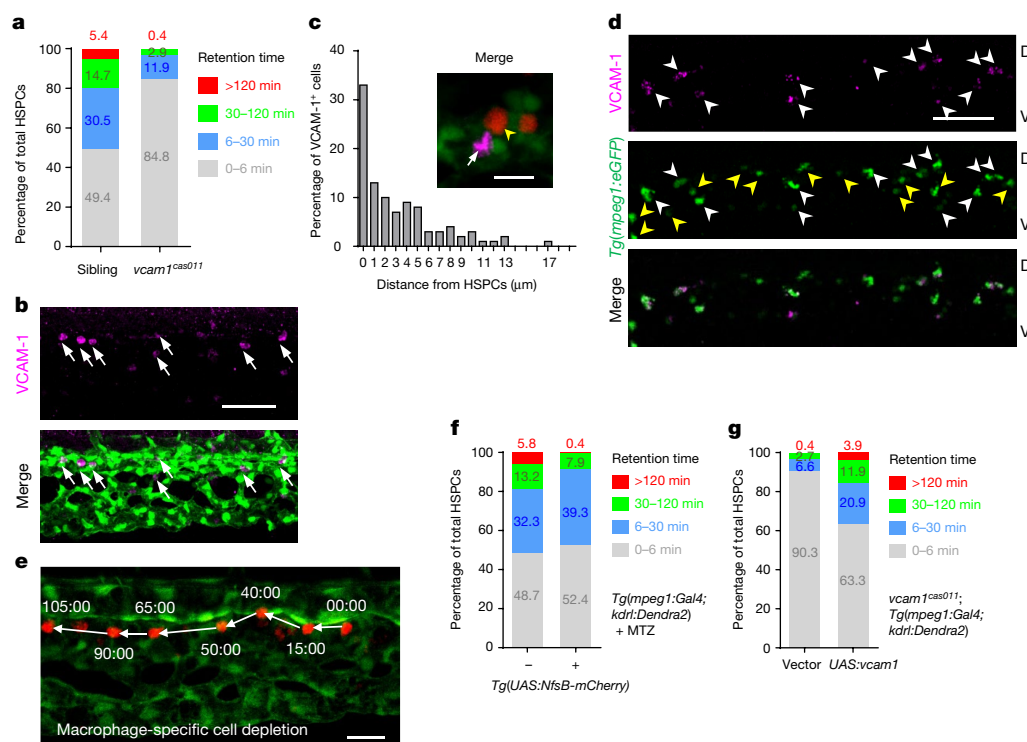


Fig. 2 | Distinct role of macrophages and venous endothelium VCAM-1 in HSPCs retention. **a**, Percentage of total HSPCs in four classified retention time zones in grouped wild-type siblings and *vcam1^{cas011}* mutants ($n = 3$) at 50–60 h.p.f. **b**, *Tg(kdrl:eGFP)* embryos, stained with an anti-VCAM-1 antibody (magenta, arrows), show dorsal venous plexus distribution of individual VCAM-1⁺ cells in the CHT. **c**, Percentage of VCAM-1⁺ cells in the CHT scored by the distance to the nearest HSPC (edge to edge, $n = 100$). Most (86%; 86 out of 100 cells) VCAM-1⁺ cells were located within 7 μm of HSPCs (the average diameter of HSPCs is about 6.9 μm). *Tg(kdrl:Dendra2)* embryos with AGM photoconversion were stained with an anti-VCAM-1 antibody (magenta, white arrow). Yellow arrowheads denote HSPCs. **d**, The staining of *Tg(mpeg1:eGFP)* embryos with an anti-VCAM-1 antibody (magenta)

shows that VCAM-1⁺ cells merge with *mpeg1*⁺ cells (green) in the CHT. White arrowheads denote VCAM-1⁺GFP⁺ double-positive cells; yellow arrowheads denote GFP single-positive cells. **e**, Live-imaging frame shots of HSPCs in macrophage-specific cell-depletion embryos from **f**. See Supplementary Video 6. Time is in minutes:seconds. **f**, Percentage of total HSPCs in four classified retention time zones in grouped ($n = 3$) *Tg(mpeg1:Gal4; kdrl:Dendra2)* embryos with MTZ treatment, with or without the *Tg(UAS:NfsB-mCherry)* background, at 50–60 h.p.f. **g**, Percentage of total HSPCs in four classified retention time zones in grouped ($n = 3$) *vcam1^{cas011}* mutants in a *Tg(mpeg1:Gal4; kdrl:Dendra2)* background with transient transgenesis of either vector (*UAS:polyA*) or *UAS:vcam1* at 50–60 h.p.f. See Supplementary Video 6. Scale bars, 50 μm (**b**, **d**), 20 μm (**e**) and 10 μm (**c**).

be needed for HSPCs to enter the venous capillaries in an ITGA4-dependent manner.

Vascular cell adhesion molecule-1 (VCAM-1) is known as the major ligand for VLA-4 in mammalian cells^{21,22}. According to the ZFIN database (<http://zfin.org/ZDB-GENE-070209-238>), zebrafish *vcam1* (also known as *vcam1b*) was specifically expressed in the cranial region, heart and the CHT at around 30 h.p.f. To evaluate the function of VCAM-1 in definitive haematopoiesis, we generated a *vcam1^{cas011}* mutant (Extended Data Fig. 7a, b). The mutants resembled the defects in homing and definitive haematopoiesis observed in the *itga4* mutants (Fig. 2a, Extended Data Figs. 2f, g, 7c–f, Supplementary Video 5), indicating that the ITGA4–VCAM-1 axis has an evolutionarily conserved role in nascent HSPC homing and retention^{16,23}.

Immunofluorescence staining showed that endogenous VCAM-1 was strongly expressed on cells that are mostly distributed at the dorsal CVP, where HSPCs show preferential lodgement (Fig. 2b). In addition, VCAM-1 protein was also weakly expressed on some of the venous endothelial cells in the CHT (Extended Data Fig. 7g). Importantly, the VCAM-1⁺ non-endothelial cells in the CHT were always next to HSPCs (Fig. 2c, Extended Data Fig. 7h), which were neither previously described *cxcl12a:DsRed*⁺ cells nor somite-derived stromal reticular cells^{8,9} (Extended Data Fig. 7i, j). By contrast, we found that almost all the VCAM-1⁺ non-endothelial cells were GFP⁺ in the macrophage-specific *Tg(mpeg1:eGFP)* transgenic line in the CHT (Fig. 2d). Meanwhile, about 45% of *mpeg1*⁺ cells in the CHT are VCAM-1⁺, and there are on average 13 VCAM-1⁺ macrophage-like

cells per CHT (Extended Data Fig. 8d) that express the macrophage markers *mfap4*, *csf1ra* and *spi1a* with high overlapping rates (Extended Data Fig. 8a and Supplementary Table 2). Thus, these VCAM-1⁺ non-endothelial cells in the CHT are likely to be a subtype of macrophages.

To characterize the potential function of these VCAM-1⁺ macrophage-like cells in the homing and retention of HSPCs, we either depleted macrophages using metronidazole (MTZ) (loss-of-function analysis; Extended Data Fig. 8b–f), or transiently expressed wild-type *vcam1* in *mpeg1*-positive macrophage cells in *vcam1* mutants (gain-of-function analysis; Extended Data Fig. 8b–g). MTZ depletion of macrophages did not affect HSPC emergence (Extended Data Fig. 8c), but caused impaired HSPC lodgement (Fig. 2e, f, Extended Data Fig. 8f) and defective definitive haematopoiesis (Extended Data Fig. 8e), indicating that macrophages are essential for HSPC retention. However, when the behaviour of HSPCs was compared with those of the *vcam1^{cas011}* mutants, we found that although HSPCs in macrophage-depleted embryos did not successfully lodge in the CHT, they could flow slowly in the vasculature, suggesting that endothelial VCAM-1 might have a role in the initiation of HSPC rolling on the dorsal endothelium bed (Fig. 2e, Extended Data Figs. 4e, 8f and Supplementary Video 6), consistent with previous reports²⁴.

The re-expression of *vcam1* exclusively in *mpeg1*-positive cells could significantly restore HSPC retention (Fig. 2g, Extended Data Figs. 4e, 8e–g and Supplementary Video 6). The incomplete rescue of the retention phenotype suggests that the interaction of HSPCs with the CVP

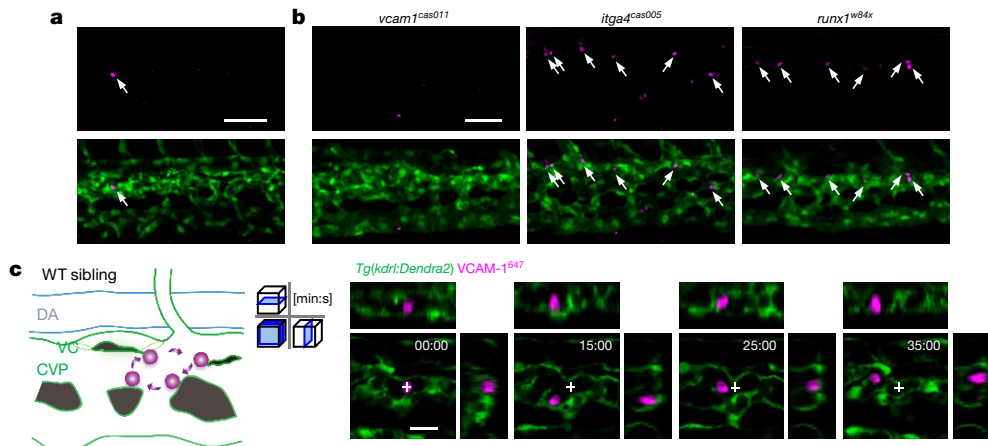


Fig. 3 | Characterization of VCAM-1⁺ macrophages in the CHT.

a, Transgenic *Tg(kdr:eGFP)* embryos, stained with an anti-VCAM-1 antibody (magenta, arrows), show that the VCAM-1⁺ macrophage first appeared in the CHT at 32 h.p.f. **b**, *Tg(kdr:eGFP)* embryos in the *vcam1*^{cas011}, *itga4*^{cas005} or *runx1*^{w84x} mutant background are stained with an anti-VCAM-1 antibody (magenta, white arrows) at 54 h.p.f. Signals in *itga4*^{cas005} and *runx1*^{w84x} are similar to that in wild-type siblings, whereas there is almost no detectable signal in *vcam1*^{cas011} mutants. **c**, Schematic diagrams (left) and confocal

imaging (right) of VCAM-1⁺ macrophages (labelled with Alexa Fluor 647 dye-conjugated anti-VCAM-1 antibody by intravascular injection) that patrol the CHT of wild-type embryos. VCAM-1⁺ macrophages were mainly located intravascularly (>91%) with round or unpolarized cell morphology (>84%). Cross indicates the original position of VCAM-1⁺ macrophages at the initial imaging time point. See Supplementary Video 7. DA, dorsal aorta; VC, venous capillaries. Scale bars, 50 μ m (**a**, **b**) and 20 μ m (**c**).

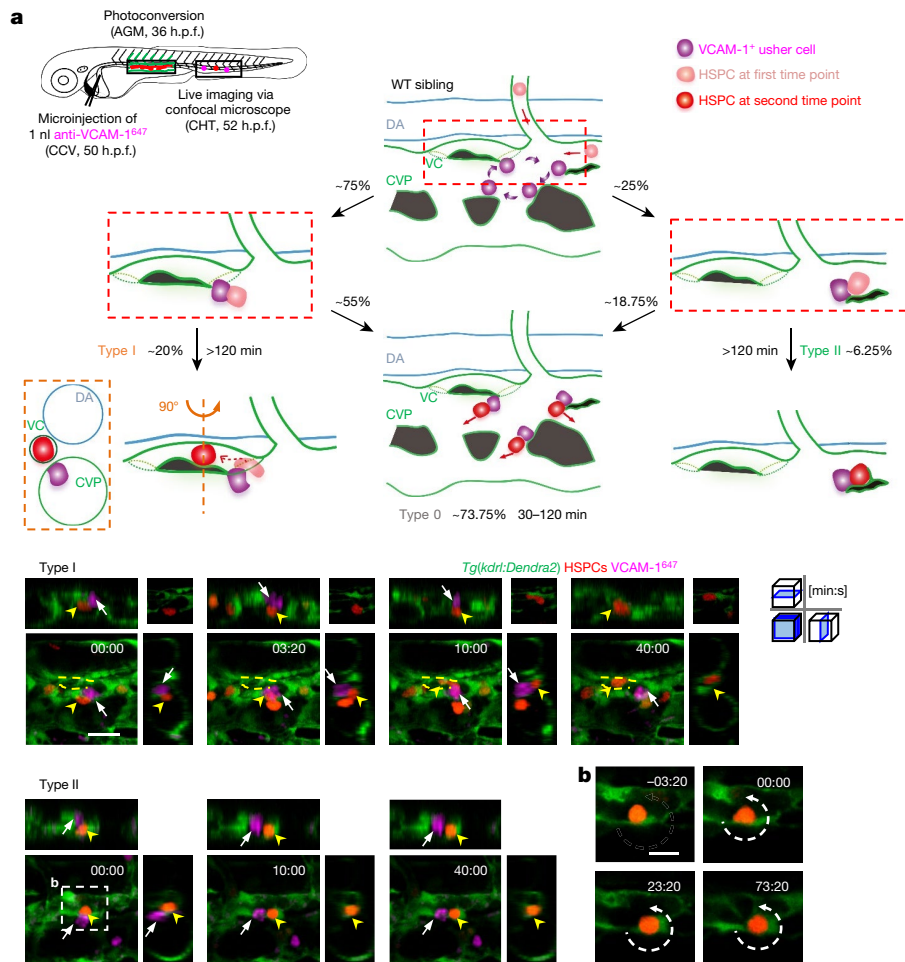


Fig. 4 | Live imaging analysis on VCAM-1⁺ usher cell-guided HSPCs retention. **a**, Schematic illustration (top left) shows that labelling of HSPCs (with photoconverted Dendra2, red) was performed at 36 h.p.f. in *Tg(kdr:Dendra2)* embryos, followed by an anti-VCAM-1⁶⁴⁷ antibody injection at 50 h.p.f. Live imaging was performed 2 h after injection (52 h.p.f.). Schematic diagrams (top) show the HSPC retention model, with accurate percentage and classification. See Supplementary Table 3. Representative images (bottom) show the interaction between HSPCs (red;

yellow arrowheads) and VCAM-1⁺ usher cells (magenta; white arrows). VCAM-1⁺ usher cells guide HSPCs into the venous capillaries, leading to type I and II retention (see Supplementary Videos 8 and 9). The top right images show one slice, and the others show z-stacks. **b**, Enlarged view of endothelial cell remodelling around a single HSPC to form a stem-cell pocket in type II retention (see Supplementary Video 9). Scale bars, 20 μ m (**a**) and 10 μ m (**b**).

endothelium might also have a role by slowing down HSPCs to increase the chance for HSPC retention, even though the HSPC–CVP endothelium interaction is not absolutely required for HSPC retention.

To determine when the VCAM-1⁺ macrophages appear in the CHT, we performed a time-course analysis of VCAM-1 expression from 28 to 48 h.p.f. The immunofluorescence results demonstrated that the VCAM-1⁺ macrophages first appeared in the CHT at 32 h.p.f. (Fig. 3a). The VCAM-1⁺ macrophages in the CHT were absent in the *vcam1*^{cas011} mutants and independent of HSPC deficiency in either *itga4*^{cas005} or *runx1*^{w84x} mutants at 54 h.p.f. (Fig. 3b). Because the VCAM-1⁺ macrophages are present in the niche before the appearance of aorta-derived definitive HSPCs, these macrophages probably arise from the primitive macrophage lineage at this time point. We labelled the primitive macrophages by applying photoconversion on the *Tg(mpeg1:Gal4,UAS:Kaede)* line, and found that some macrophages from the rostral blood island²⁵ at 18 h.p.f. migrated to the CHT, and were VCAM-1⁺ (Extended Data Fig. 8h).

To characterize the behaviour and function of the VCAM-1⁺ macrophages in live animals further, we labelled VCAM-1⁺ macrophages with an anti-VCAM-1⁶⁴⁷ antibody. The live imaging showed a nearly identical cell distribution pattern to that revealed by anti-VCAM-1 immunofluorescence (Extended Data Fig. 9a). Live staining with the antibody remained stable for at least 8 h after intravascular antibody injection, without affecting definitive haematopoiesis, as demonstrated by quantitative *myb* WISH analysis (Extended Data Fig. 9b–d).

Notably, these VCAM-1⁺ macrophages slowly patrolled on the inner sides, especially the dorsal CVP (Fig. 3c, Extended Data Fig. 9e, Supplementary Video 7). HSPCs, entering from either the ISV or the CVP into the CHT, always pass by the capillary confluence point, which leads to frequent interactions of HSPCs with the VCAM-1⁺ macrophages.

After quantitative analysis of more than 100 VCAM-1⁺ macrophage–HSPC interactions and subsequent events (Fig. 4a, Supplementary Table 3), we found that, on average, each interaction lasted approximately 30 min. About 60% of the HSPCs left the CHT through the venous plexus without retention (6–30 min), whereas around 40% of HSPCs could remain in the CHT for more than 30 min.

Among these lodged HSPCs, about 75% interacted with VCAM-1⁺ macrophages at the entrance of dorsal venous capillaries. With the guidance of VCAM-1⁺ macrophages, 20% could finally enter the venous capillaries and remained for more than 120 min. We defined this as ‘type I’ retention (Fig. 4a, Extended Data Fig. 9f, Supplementary Video 8). Conversely, 25% of the lodged HSPCs interacted with VCAM-1⁺ macrophages within the CVP, and then 6.25% were surrounded by an ‘endothelial pocket’¹⁴ structure, leading to the ‘type II’ retention (Fig. 4a, b, Extended Data Fig. 9f, Supplementary Video 9). HSPCs that successfully interacted with VCAM-1⁺ macrophages (for more than 30 min) but failed to be guided into the vascular niche (55% as pre-type I and 18.75% as pre-type II retention) were termed ‘type 0’ retention. Notably, HSPCs have an increased chance of interacting with VCAM-1⁺ macrophages at venous capillary confluence points connected to the CVP where dorsal venous capillaries are also distributed and the hotspots of HSPCs retention were observed (Fig. 1f).

In *itga4*^{cas010} mutants, HSPCs encountered but did not interact with VCAM-1⁺ macrophages. They could not roll inside the vasculature, enter venous capillaries or be enveloped by endothelium. Instead, the HSPCs went quickly through the CHT (Extended Data Fig. 9g and Supplementary Video 10). Thus, the interaction mediated by ITGA4 and VCAM-1 has an essential role in HSPC homing and retention in the niche. Together with the feature that VCAM-1⁺ macrophages patrol at the dorsal CVP, we named these VCAM-1⁺ macrophages ‘usher’ cells. In addition, HSPCs retention in *vcam1*^{cas011} mutants with macrophage-specific *vcam1* re-expression restored the occurrence of type II retention, indicating that the loss of vascular VCAM-1 is not sufficient to disrupt the vascular ‘cuddling’ structure (Extended Data Fig. 9h).

In this study, we mainly focused on the period of 48–60 h.p.f. at the initiation of HSPC homing because current imaging technology and transgenic lines do not allow the symmetry division to be distinguished from the asymmetry one, although most of the HSPC division in the CHT occurred in the vascular niche (66% in type I, 30% in type II retention). Morphology-based asymmetry division of HSPCs was previously reported in the vascular cuddling structure¹⁴; however, lineage-specific reporter lines with rapid responses are required in future studies to understand how the dynamic HSPC–niche interaction is coordinated with HSPC division.

We show that the behaviour of usher cells correlates with the retention of HSPCs in the CHT. It has been reported that macrophages promote the retention of HSPCs^{26,27}. However, it is still not clear how the usher cells as well as niche cells express other molecules that recognize receptors on the HSPCs, serving as additional permissive signals for the entry of HSPCs into the niches. In addition, the *itga4* mutant specifically impaired definitive haematopoiesis in the CHT, but not that in the thymus or kidney marrow, indicating different homing mechanisms might govern the lodgement of HSPCs into different niches (Extended Data Fig. 1e, g). Future studies are warranted to investigate these important questions.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0709-7>.

Received: 7 December 2017; Accepted: 14 September 2018;

Published online: 19 November 2018

- Morrison, S. J., Uchida, N. & Weissman, I. L. The biology of hematopoietic stem cells. *Annu. Rev. Cell Dev. Biol.* **11**, 35–71 (1995).
- Murayama, E. et al. Tracing hematopoietic precursor migration to successive hematopoietic organs during zebrafish development. *Immunity* **25**, 963–975 (2006).
- Kissa, K. & Herbomel, P. Blood stem cells emerge from aortic endothelium by a novel type of cell transition. *Nature* **464**, 112–115 (2010).
- Bertrand, J. Y. et al. Haematopoietic stem cells derive directly from aortic endothelium during development. *Nature* **464**, 108–111 (2010).
- Boisset, J. C. et al. In vivo imaging of haematopoietic cells emerging from the mouse aortic endothelium. *Nature* **464**, 116–120 (2010).
- Xue, Y. et al. The vascular niche regulates hematopoietic stem and progenitor cell lodgment and expansion via *klf6a-ccl25b*. *Dev. Cell* **42**, 349–362.e4 (2017).
- Mahony, C. B., Fish, R. J., Pasche, C. & Bertrand, J. Y. *ttec* controls the hematopoietic stem cell vascular niche during zebrafish embryogenesis. *Blood* **128**, 1336–1345 (2016).
- Glass, T. J. et al. Stromal cell-derived factor-1 and hematopoietic cell homing in an adult zebrafish model of hematopoietic cell transplantation. *Blood* **118**, 766–774 (2011).
- Murayama, E. et al. NACA deficiency reveals the crucial role of somite-derived stromal cells in hematopoietic niche formation. *Nat. Commun.* **6**, 8375 (2015).
- Jin, H., Xu, J. & Wen, Z. Migratory path of definitive hematopoietic stem/progenitor cells during zebrafish development. *Blood* **109**, 5208–5214 (2007).
- Burns, C. E., Traver, D., Mayhall, E., Shepard, J. L. & Zon, L. I. Hematopoietic stem cell fate is established by the Notch–Runx pathway. *Genes Dev.* **19**, 2331–2342 (2005).
- Jao, L. E., Wente, S. R. & Chen, W. Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. *Proc. Natl Acad. Sci. USA* **110**, 13904–13909 (2013).
- Soza-Ried, C., Hess, I., Netuschil, N., Schorpp, M. & Boehm, T. Essential role of *c-myc* in definitive hematopoiesis is evolutionarily conserved. *Proc. Natl Acad. Sci. USA* **107**, 17304–17308 (2010).
- Tamplin, O. J. et al. Hematopoietic stem cell arrival triggers dynamic remodeling of the perivascular niche. *Cell* **160**, 241–252 (2015).
- Arroyo, A. G., Yang, J. T., Rayburn, H. & Hynes, R. O. $\alpha 4$ integrins regulate the proliferation/differentiation balance of multilineage hematopoietic progenitors *in vivo*. *Immunity* **11**, 555–566 (1999).
- Imai, Y., Shimaoka, M. & Kurokawa, M. Essential roles of VLA-4 in the hematopoietic system. *Int. J. Hematol.* **91**, 569–575 (2010).
- Gribi, R., Hook, L., Ure, J. & Medvinsky, A. The differentiation program of embryonic definitive hematopoietic stem cells is largely $\alpha 4$ integrin independent. *Blood* **108**, 501–509 (2006).
- Qian, H. et al. Distinct roles of integrins $\alpha 6$ and $\alpha 4$ in homing of fetal liver hematopoietic stem and progenitor cells. *Blood* **110**, 2399–2407 (2007).
- Scott, L. M., Priestley, G. V. & Papayannopoulou, T. Deletion of $\alpha 4$ integrins from adult hematopoietic cells reveals roles in homeostasis, regeneration, and homing. *Mol. Cell. Biol.* **23**, 9349–9360 (2003).

20. Nombela-Arrieta, C. et al. Quantitative imaging of haematopoietic stem and progenitor cell localization and hypoxic status in the bone marrow microenvironment. *Nat. Cell Biol.* **15**, 533–543 (2013).
21. Osborn, L. et al. Direct expression cloning of vascular cell adhesion molecule 1, a cytokine-induced endothelial protein that binds to lymphocytes. *Cell* **59**, 1203–1211 (1989).
22. Elices, M. J. et al. VCAM-1 on activated endothelium interacts with the leukocyte integrin VLA-4 at a site distinct from the VLA-4/fibronectin binding site. *Cell* **60**, 577–584 (1990).
23. Koenig, J. M., Ballantyne, C. M., Kumar, A. G., Smith, C. W. & Yoder, M. C. Vascular cell adhesion molecule-1 expression and hematopoietic supportive capacity of immortalized murine stromal cell lines derived from fetal liver and adult bone marrow. *In Vitro Cell. Dev. Biol. Anim.* **38**, 538–543 (2002).
24. Berlin, C. et al. $\alpha 4$ integrins mediate lymphocyte attachment and rolling under physiologic flow. *Cell* **80**, 413–422 (1995).
25. Warg, R. M., Kane, D. A. & Ho, R. K. Fate mapping embryonic blood in zebrafish: multi- and unipotential lineages are segregated at gastrulation. *Dev. Cell* **16**, 744–755 (2009).
26. Winkler, I. G. et al. Bone marrow macrophages maintain hematopoietic stem cell (HSC) niches and their depletion mobilizes HSCs. *Blood* **116**, 4815–4828 (2010).
27. Dutta, P. et al. Macrophages retain hematopoietic stem cells in the spleen via VCAM-1. *J. Exp. Med.* **212**, 497–512 (2015).

Acknowledgements We thank the following people for the zebrafish transgenic lines: L. Luo for *Tg(kdrl:Dendra2)*, Z. Wen for *Tg(mpeg1:Gal4,UAS:NfsB-mCherry)*, *Tg(UAS:Kaede)* and *Tg(mpeg1:eGFP)*, B. Blazar for *Tg(cxcl12a:dsRed)* and F. Argenton for *Tg(tcf:eGFP)*. We are also grateful to M. Deng and J. He for technical support, and Z. Wen, L. Li, L. Zon, J. Peng and A. Meng for discussions. This work was granted by CAS Strategic Priority Research Program (XDB19030000), Ministry of Science and Technology of China (2017YF0503600), National Natural Science Foundation of China (31571505

and 31371461), CAS Scientific Research Equipment Development Project (YZ201646) and Science and Technology Commission of Shanghai Municipality (13JC1406400) to W.J.P.

Reviewer information *Nature* thanks P. Herbolme and the anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions D.L., W.X. and W.P. developed the concepts and designed the experiments. D.L., W.X. and M.L. performed the experiments and analysed data. M.D. performed ENU screening and positional cloning. J.W. assisted with 3D reconstruction and HSPC retention heat-map analysis. X.W. and X.L. assisted the experiments and data analysis during revision. K.C., W.Z. and S.W. assisted with the schematic illustration of the working model for HSPC homing. Y.J.C. assisted with imaging with Zeiss 880 and Y. Zhang assisted with 3D-reconstruction analysis. J.C. and X.Y. supported experiments and provided ideas about the distinct role of VCAM-1 in different cell populations. L.G., B.O.Z., Y. Zhou, L.L. and D.W. provided ideas and discussions throughout the project. J.C., B.O.Z., Y. Zhou, X.Y. and L.L. revised the manuscript. D.L., D.W. and W.P. wrote the paper. W.P. supervised the project.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0709-7>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0709-7>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to W.P.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Zebrafish husbandry. The zebrafish facility and study were approved by the Animal Research Advisory Committee of Institute of Nutrition and Health, SIBS, CAS, and zebrafish were maintained according to the guidelines of the Institutional Animal Care and Use Committee. The Tubingen and WIK wild-type strains were used in this study. The *runx1*^{w84x} mutant line²⁸ and transgenic lines *Tg(gata1:DsRed)*²⁹, *Tg(mpx:eGFP)*³⁰, *Tg(lyz:DsRed)*³¹, *Tg(kdrl:Dendra2)*³², *Tg(kdrl:eGFP)*³³, *Tg(mpeg1:Gal4,UAS:NfsB-mCherry)*³⁴, *Tg(UAS:Kaede)*³⁵, *Tg(cxcl12a:DsRed)*⁸, *Tg(tcf:eGFP)*^{9,36} and *Tg(mpeg1:eGFP)*³⁷ were described previously.

Genetic mapping and mutation identification of zebrafish *cas005* mutant. The *mut*⁰⁰⁵ line was identified in a large-scale *N*-ethyl-*N*-nitrosourea (ENU) -mutagenized F₂ family screen based on the *myb* expression phenotype of 5 days post-fertilization (d.p.f.) zebrafish embryos. The ENU screen and positional cloning were performed as described previously³⁸. The mutation was mapped to chromosome 9 by bulk segregation analysis with sequence length polymorphism (SSLP) markers³⁹. Fine mapping was carried out to narrow down the genetic interval, and the mutation was finally flanked by two SSLP markers, z8363 and zK165L22. The cDNAs of candidate genes in the range were cloned and sequenced from siblings and mutants, and the putative mutation was confirmed by sequencing genomic DNA of individual mutant embryos. All primers used for this study are provided in Supplementary Table 1.

Plasmid construction. The zebrafish *vcam1* (also known as *vcam1b*, accession number: ZDB-GENE-070209-238) was cloned and inserted into the Tol2 backbone between the UAS promoter and polyA. The *runx1* +23 enhancer followed by P2A and in-frame mCherry was cloned into the Tol2 backbone, and the zebrafish *itga4* (accession number: ZDB-GENE-110411-108) was amplified and inserted between the *runx1* enhancer and P2A. The sequences of *vcam1* and *itga4* were verified by sequencing. The *runx1* +23 enhancer¹⁴ and UAS promoter⁴⁰ were cloned as previously reported.

Microinjection and CRISPR-Cas9 mutagenesis. Morpholino oligonucleotides were designed and purchased from Gene Tools. The morpholino oligonucleotides used in this study give the same phenotypes as mutants. The *itga4*, *runx1*⁴¹, *myb*¹³ and *vcam1* morpholino oligonucleotides were injected into one-cell-stage embryos as previously described⁴². Transient transgenic constructs within Tol2 vectors (40 pg) were microinjected into one-cell-stage embryos with Tol2 transposase mRNA (40 pg)⁴³. For CRISPR-Cas9-mediated generation of zebrafish mutants (*itga4*^{cas010} and *vcam1*^{cas011}), guide RNAs (gRNAs) were designed to target genes according to methods previously described⁴⁴. The zebrafish codon-optimized Cas9 mRNA was synthesized from the pCS2-nCas9n plasmid (Addgene, plasmid47929)¹² and gRNAs were in vitro synthesized using the MAXscript T7 kit (Ambion). The gRNAs (100 pg) were microinjected into one-cell-stage embryos with Cas9 mRNA (300 pg).

Conventional WISH, FISH and RNA scope in situ analysis. The *myb*⁴⁵, *scf*⁴⁶, *gata1*⁴⁶, *pu.1*⁴⁶, *kdrl*⁴⁵, *runx1*⁴⁷, *hbae1.1*⁴⁵, *mpx*⁴⁵, *lyz*⁴⁵, *itga4*, *mfap4*⁴⁸ and *csflra*⁴⁸ probes were transcribed in vitro by T3 or T7 polymerase (Ambion) with Digoxigenin RNA Labelling Mix (Roche). Conventional and fluorescence whole-mount in situ hybridization (WISH and FISH) was described previously⁴⁹. The penetrance of the indicated phenotype is shown in the bottom right of each panel (Extended Data Figs. 1a–e, g, 2d, e, h, 7d–f and 9d) and the embryos were genotyped after WISH and before phenotypic analysis (Extended Data Figs. 2d, j, k, 7d–f, 8e and 9b). Images of conventional WISH were mounted in 4% methylcellulose and captured by Olympus SZX16 microscope with Olympus DP80 CCD. In FISH and immunofluorescence double-staining, embryos were stained with cy3 or cy5 (TSA system, Perkin Elmer), followed by immunofluorescence, and then imaged by Zeiss LSM880 confocal microscope. RNA scope was conducted with probe *runx1* (P/N: 433351, ACDBio) and negative control probe (REF: 320871, ACDBio). RNA scope procedure was performed as previously described⁵⁰, and imaged by Olympus FV1000 Fluoview scanning confocal microscope. A list of oligonucleotides used to amplify these probes is provided in Supplementary Table 1.

Immunofluorescence staining and usher cell live immunolabelling. Immunofluorescence staining was performed as previously described⁵¹, with mouse anti-DsRed (Abcam), mouse anti-eGFP (Abmart), rabbit anti-VCAM-1 antibody (immunized by 180–300 amino acids of zebrafish VCAM-1 protein, Abclonal), AF488/546/647-conjugated secondary antibody (Life Technologies) and TUNEL assay kit (Roche). Images were collected using Olympus FV1000 Fluoview scanning confocal, Zeiss LSM710 confocal and Zeiss LSM880 confocal microscope and the embryos were genotyped following imaging analysis (Fig. 3b and Extended Data Fig. 8d). For live labelling of usher cells, VCAM-1 antibody (Abclonal) was conjugated with Alexa Fluor 647 dye and purified by Microscale Protein Labelling Kit (Invitrogen, A30009). Each embryo was injected 1 nl (0.4 ng) at the common cardinal vein into the circulation at 50 h.p.f. In vivo monitoring started from 2 h after injection and this live labelling is stable for more than 8 h⁵². Time-lapse intravital imaging was acquired by Zeiss LSM880 confocal microscope.

Inducible macrophage-specific cell depletion. As previously reported, MTZ-mediated cell depletion was performed on *Tg(mpeg1:GAL4,UAS:NfsB-mCherry, kdrl:Dendra2)*^{34,53} transgenic zebrafish embryos. Embryos from 24 to 60 h.p.f. were treated with freshly prepared 10 mM MTZ (Sigma) in 0.2% DMSO (dimethylsulfoxide) solution protected from light until the evaluation finished, then rinsed with embryo water three times.

Confocal microscope photoconversion and time-lapse live imaging analysis.

The photoconversion of irreversible monomeric green-to-red fluorescent protein Dendra2^{54,55} expressed following specific promoter was conducted with a 405-nm laser for 30 s by an Olympus FV1000 Fluoview scanning confocal microscope. For the HSPC labelling system, the ventral endothelium of dorsal aorta (between somites 8 and 17) of *Tg(kdrl:Dendra2)* embryos was exposed to a beam of 405 nm ultraviolet (UV) laser light under a confocal microscope at 30–36 h.p.f. without affecting the normal endothelial-to-haematopoietic transition process, compared with that in untreated embryos⁴. The efficiency of cell labelling was confirmed under fluorescent microscope 8 h after photoconversion. The zebrafish with precise and bright photoconverted Dendra2 (red) cells were selected for further analysis. For the macrophage labelling system, the rostral blood island of *Tg(mpeg1:Gal4,UAS:Kaede)* zebrafish embryos was photoconverted at 18 h.p.f., followed by anti-VCAM-1⁶⁴⁷ injection at 50 h.p.f., and imaging at 52 h.p.f. using Zeiss LSM880 confocal microscope.

For time-lapse imaging, embryos were anaesthetized with 0.03% Tricaine (Sigma), and mounted in 1.5% low melting point agarose in a 60-mm dish. The embryos were scanned at 28.5 °C under an Olympus FV1000 Fluoview scanning confocal microscope with 20× water immersion objective, z-stacks were acquired with a step size of 3 μm within an interval of 3 min over several hours. To observe more details between HSPCs and vascular niche, images were collected using a Zeiss LSM880 confocal by 40× water immersion objective; z-stacks were acquired with a size of 3 μm for over 8 h. The Zeiss LSM880 confocal allowed imaging of several embryos within a 2–5-min interval using a moving XY stage, as well as acquisition of z-stacks through the tissue in multiple channels. Note that 10 h is the maximum experimental duration of live-imaging analysis without phototoxicity. The embryos were genotyped following live-imaging analysis (Fig. 1a, c–e, 2a, e–g and Extended Data Figs. 3e, 4d, e, 6g, h, 7c, 8f, g, 9a, e, g, h).

Imaging data processing and rendering was performed in FV10-ASW 3.0 Viewer (Olympus), ZEN 2.1 (ZEISS), Imaris (Bitplane) and ImageJ (NIH). The retention time and location information that each HSPC appears in the CHT was exported by Imaris 'Spots' module and programmed by Python into retention heatmap (<https://pypi.python.org/pypi/pyheatmap>). The velocity of photoconverted HSPCs in the CHT was measured with the axial line scanning (ALS) as previously described^{56–58}. Zeiss LSM880 equipped with Airyscan function was applied to capture high-resolution fluorescent images, followed with 3D reconstruction by Materialise software (including Mimics Medical and 3-matic Medical).

Statistical analysis. All statistical analysis was performed using Graphpad Prism 7 software using the two-tailed Student's *t*-test. Centre values denote the mean, and error values denote s.e.m. (Fig. 1c and Extended Data Figs. 1f, 2f, 3d, 6h, 8c, d, 9f). The biologically independent sample size (*n*) was shown in the relevant figure panel (Fig. 1c and Extended Data Figs. 1f, 2f, j, 3d, 6h, 8c–e, 9b, f, h). All experiments in this study were repeated independently at least three times. For representative images, we have performed imaging on 5–10 embryos per independent experiments and repeated at least three times independently to find the most representative images (Figs. 1a, 2b–e, 3a–c, 4 and Extended Data Figs. 2k, 3e, 5a, c, d, 6e–g, 7c, g–j, 8a, c, g, h, 9a, c, e, g). The graphs in Fig. 1d and Extended Data Fig. 8f show individual values in three embryos per group separately. In Fig. 1f and Extended Data Fig. 4b–e, imaging was performed on one embryo per independent experiment, repeated three times independently, and three images were chosen for the analysis. *****P* < 0.0001. No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

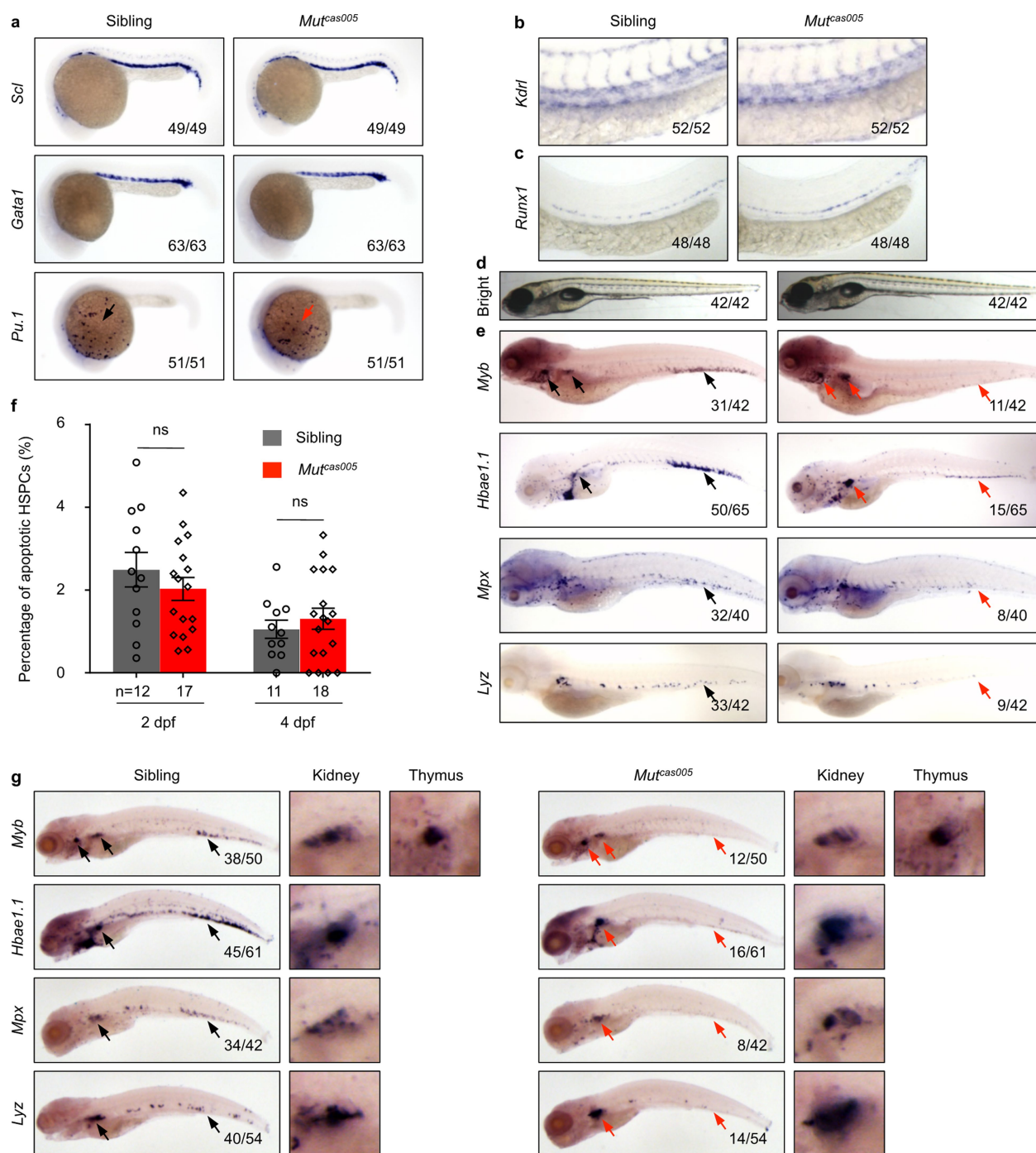
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper. 3D reconstruction of vessel surrounding HSPCs in retention hotspot is deposited at <http://www.biosino.org/node/project/detail/OEP000169>.

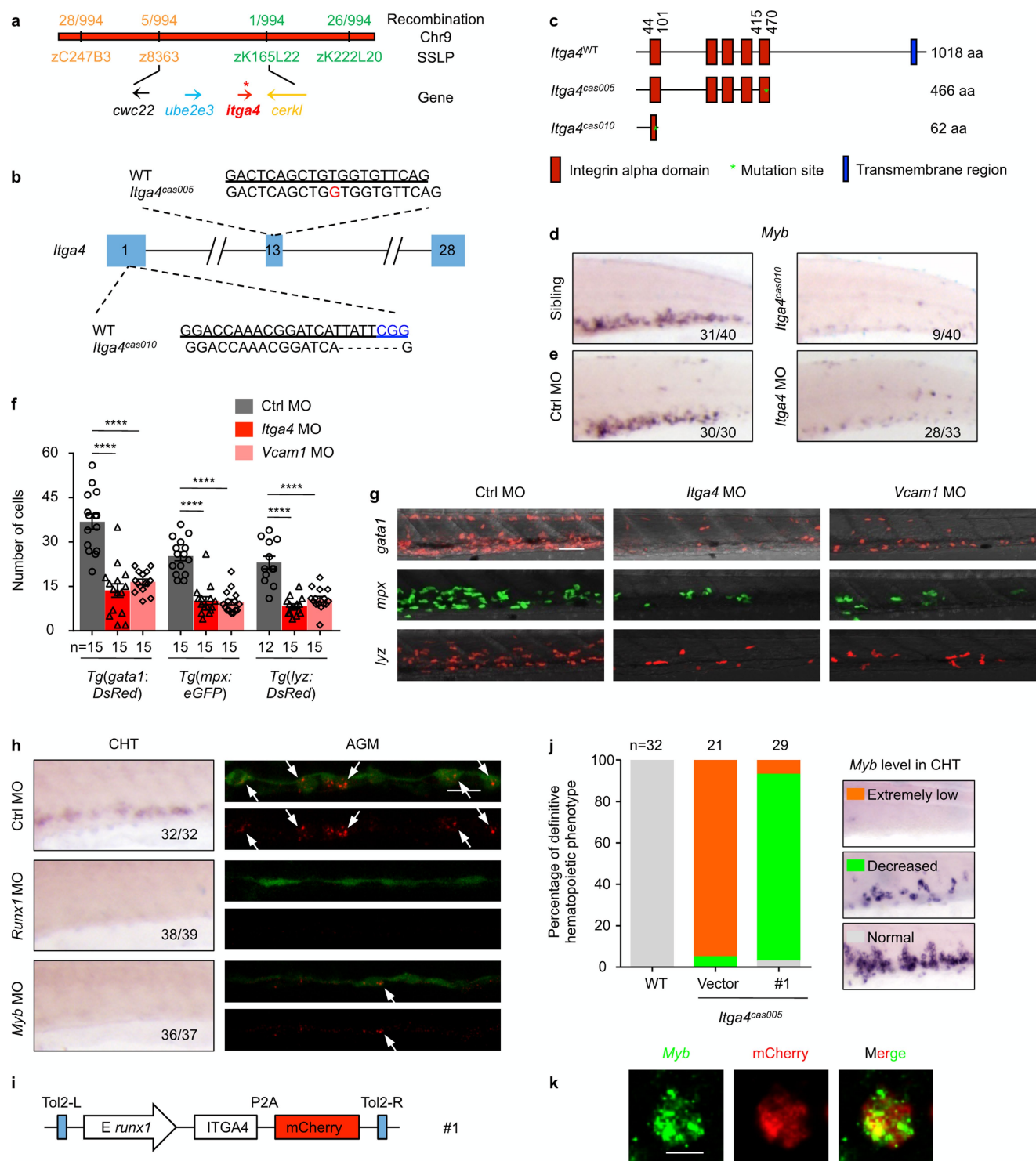
28. Jin, H. et al. Definitive hematopoietic stem/progenitor cells manifest distinct differentiation output in the zebrafish VDA and PBL. *Development* **136**, 647–654 (2009).
29. Traver, D. et al. Transplantation and in vivo imaging of multilineage engraftment in zebrafish bloodless mutants. *Nat. Immunol.* **4**, 1238–1246 (2003).

30. Shi, X. et al. Functions of *idh1* and its mutation in the regulation of developmental hematopoiesis in zebrafish. *Blood* **125**, 2974–2984 (2015).
31. Hall, C., Flores, M. V., Storm, T., Crosier, K. & Crosier, P. The zebrafish lysozyme C promoter drives myeloid-specific expression in transgenic fish. *BMC Dev. Biol.* **7**, 42 (2007).
32. Liu, C. et al. Macrophages mediate the repair of brain vascular rupture through direct physical adhesion and mechanical traction. *Immunity* **44**, 1162–1176 (2016).
33. Cross, L. M., Cook, M. A., Lin, S., Chen, J. N. & Rubinstein, A. L. Rapid analysis of angiogenesis drugs in a live fluorescent zebrafish assay. *Arterioscler. Thromb. Vasc. Biol.* **23**, 911–912 (2003).
34. Davison, J. M. et al. Transactivation from Gal4-VP16 transgenic insertions for tissue-specific cell labeling and ablation in zebrafish. *Dev. Biol.* **304**, 811–824 (2007).
35. Scott, E. K. & Baier, H. The cellular architecture of the larval zebrafish tectum, as revealed by Gal4 enhancer trap lines. *Front. Neural Circuits* **3**, 13 (2009).
36. Moro, E. et al. *In vivo* Wnt signaling tracing through a transgenic biosensor fish reveals novel activity domains. *Dev. Biol.* **366**, 327–340 (2012).
37. Ellett, F., Pase, L., Hayman, J. W., Andrianopoulos, A. & Lieschke, G. J. *mpeg1* promoter transgenes direct macrophage-lineage expression in zebrafish. *Blood* **117**, e49–e56 (2011).
38. Bahary, N. et al. The Zon laboratory guide to positional cloning in zebrafish. *Methods Cell Biol.* **77**, 305–329 (2004).
39. Knapik, E. W. et al. A microsatellite genetic linkage map for zebrafish (*Danio rerio*). *Nat. Genet.* **18**, 338–343 (1998).
40. Scheer, N. & Campos-Ortega, J. A. Use of the Gal4-UAS technique for targeted gene expression in the zebrafish. *Mech. Dev.* **80**, 153–158 (1999).
41. Lam, E. Y. N. et al. Zebrafish *runx1* promoter-EGFP transgenics mark discrete sites of definitive blood progenitors. *Blood* **113**, 1241–1249 (2009).
42. Nasevicius, A. & Ekker, S. C. Effective targeted gene ‘knockdown’ in zebrafish. *Nat. Genet.* **26**, 216–220 (2000).
43. Suster, M. L., Kikuta, H., Urasaki, A., Asakawa, K. & Kawakami, K. Transgenesis in zebrafish with the *Tol2* transposon system. *Methods Mol. Biol.* **561**, 41–63 (2009).
44. Xiao, A. et al. Chromosomal deletions and inversions mediated by TALENs and CRISPR/Cas in zebrafish. *Nucleic Acids Res.* **41**, e141 (2013).
45. Gao, L. et al. TopBP1 governs hematopoietic stem/progenitor cells survival in zebrafish definitive hematopoiesis. *PLoS Genet.* **11**, e1005346 (2015).
46. Patterson, L. J. et al. The transcription factors *Scl* and *Lmo2* act together during development of the hemangioblast in zebrafish. *Blood* **109**, 2389–2398 (2007).
47. Jia, X. E. et al. Mutation of *kri1l* causes definitive hematopoiesis failure via PERK-dependent excessive autophagy induction. *Cell Res.* **25**, 946–962 (2015).
48. Zakrzewska, A. et al. Macrophage-specific gene functions in *Spi1*-directed innate immunity. *Blood* **116**, e1–e11 (2010).
49. Jowett, T. & Lettice, L. Whole-mount in situ hybridizations on zebrafish embryos using a mixture of digoxigenin- and fluorescein-labelled probes. *TIG* **10**, 73–74 (1994).
50. Wang, H. et al. Dual-color ultrasensitive bright-field RNA in situ hybridization with RNAscope. *Methods Mol. Biol.* **1211**, 139–149 (2014).
51. Murphey, R. D., Stern, H. M., Straub, C. T. & Zon, L. I. A chemical genetic screen for cell cycle inhibitors in zebrafish embryos. *Chem. Biol. Drug Des.* **68**, 213–219 (2006).
52. Mazzocco, C. et al. *In vivo* imaging of prostate cancer using an anti-PSMA scFv fragment as a probe. *Sci. Rep.* **6**, 23314 (2016).
53. Curado, S. et al. Conditional targeted cell ablation in zebrafish: a new tool for regeneration studies. *Dev. Dyn.* **236**, 1025–1035 (2007).
54. Chudakov, D. M., Lukyanov, S. & Lukyanov, K. A. Using photoactivatable fluorescent protein Dendra2 to track protein movement. *Biotechniques* **42**, <https://doi.org/10.2144/000112470> (2007).
55. Chudakov, D. M., Lukyanov, S. & Lukyanov, K. A. Tracking intracellular protein movements using photoswitchable fluorescent proteins PS-CFP2 and Dendra2. *Nat. Protocols* **2**, 2024–2032 (2007).
56. Chen, Q. et al. Haemodynamics-driven developmental pruning of brain vasculature in zebrafish. *PLoS Biol.* **10**, e1001374 (2012).
57. Kamoun, W. S. et al. Simultaneous measurement of RBC velocity, flux, hematocrit and shear rate in vascular networks. *Nat. Methods* **7**, 655–660 (2010).
58. Schaffer, C. B. et al. Two-photon imaging of cortical surface microvessels reveals a robust redistribution in blood flow after vascular occlusion. *PLoS Biol.* **4**, e22 (2006).



Extended Data Fig. 1 | Phenotype characterization of zebrafish *mutant^{cas005}*. **a**, Normal primitive haematopoiesis is intact in *mut^{cas005}*. WISH results demonstrate that the expression of primitive haematopoietic cell markers is identical between siblings and *mut^{cas005}* embryos at 22 h.p.f., including *scl* (also known as *tall1*; haematopoietic progenitor marker), *gata1* (also known as *gata1a*; erythrocyte progenitor marker) and *pu.1* (also known as *spi1b*; myeloid progenitor marker). **b**, The vascular development is normal in *mut^{cas005}* embryos. WISH results show no difference in the expression of *kdr1* (pan-endothelial cell marker) at 36 h.p.f. between wild-type siblings and *mut^{cas005}* embryos. **c**, The haemogenic endothelium is intact in *mut^{cas005}* embryos. WISH results show no difference in *runx1* expression at 36 h.p.f. in wild-type and *mut^{cas005}* embryos. **d**, **e**, The definitive haematopoiesis is defective in zebrafish *mut^{cas005}* embryos. **d**, Bright-field images of wild-type and *mut^{cas005}* embryos show no obvious morphological difference at 5 d.p.f. **e**, WISH results of *myb*, *hbae1.1*, *mpx* and *lyz* expression in wild-type and *mut^{cas005}* embryos at 5 d.p.f. Arrows indicate the comparable position in

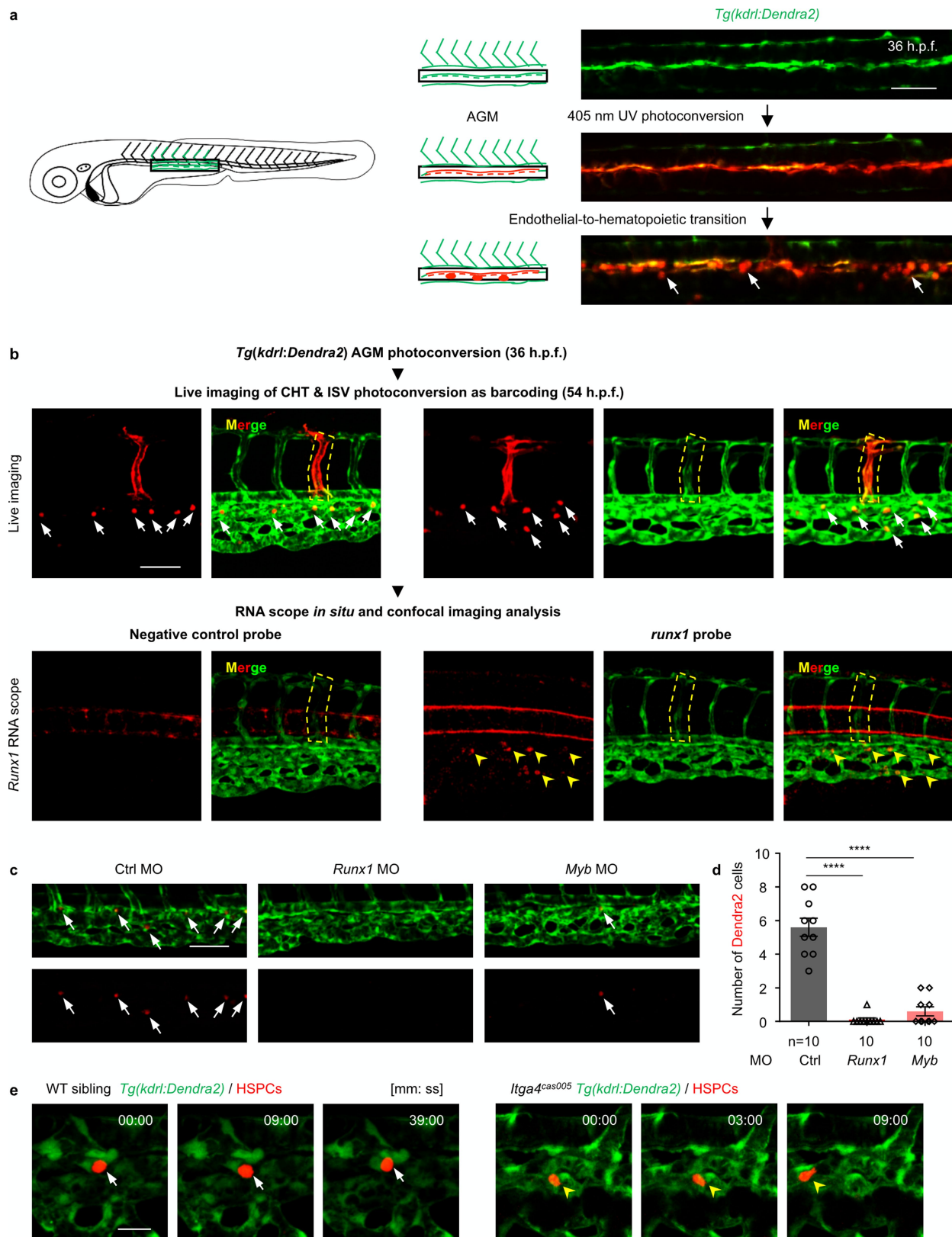
wild-type (black) or *mut^{cas005}* embryos (red). In 5 d.p.f. wild-type embryos, *myb* was expressed in all haematopoietic tissues including the CHT, thymus and kidney, whereas homozygous *mut^{cas005}* embryos displayed markedly decreased *myb* expression in the CHT, but similar expression to wild-type embryos in the thymus and kidney marrow. In accordance, the expression of downstream haematopoietic lineage cell markers, including *hbae1.1* (erythrocyte marker), *mpx* (granulocyte marker) and *lyz* (macrophage marker), also showed similar expression patterns in the wild-type (black) and *mut^{cas005}* (red) embryos to that of the *myb* WISH analysis. **f**, The percentage of apoptotic HSPCs detected by the TUNEL assay in 2 and 4 d.p.f. wild-type and *mut^{cas005}* embryos. 2 d.p.f.: $P = 0.35$, $t = 0.96$, $df = 27$; 4 d.p.f.: $P = 0.50$, $t = 0.69$, $df = 27$. Error bars denote s.e.m. **g**, At 7 d.p.f., *myb* expression in the thymus and kidney marrow was identical in wild-type sibling and *mut^{cas005}* embryos, whereas *mut^{cas005}* embryos still displayed markedly decreased *myb* expression in the CHT. In addition, *mut^{cas005}* embryos displayed notably decreased *hbae1.1*, *mpx* and *lyz* expression in the CHT but not in the kidney marrow.



Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Genetic mapping and verification of zebrafish *itga4* mutants. **a**, Positional cloning of *mut^{cas005}*. After high-resolution mapping, the point mutation is flanked by the SSLP markers z8363 (5 recombinant out of 994 meioses) and zK165L22 (1 recombinant out of 994 meioses). This region contains the four genes *cwc22*, *ube2e3*, *itga4* and *cerkl*. The red strip represents chromosome 9; the positions and recombination of the SSLP markers are indicated. SSLP markers on the same side of the mutation site are shown in the same colour. **b**, Generation of *itga4* mutants via the ENU (top) or CRISPR–Cas9 (bottom) technique. The alignment of wild-type (underlined) and mutated sequences is listed. The insertion in ENU is indicated in red (an insertion of G leading to an earlier stop codon in the *itga4* gene in *mut^{cas005}*). The PAM sequence of gRNA is ‘CGG’ (blue). Deletions are indicated by dashes. **c**, According to the stop codon in the genome of *itga4* mutants, SMART software was used to predict the structure of the wild-type *itga4*, *itga4^{cas005}* and *itga4^{cas010}* presumed protein. The molecular sizes of the presumed protein are indicated. **d**, The *myb* WISH results in wild-type and *itga4^{cas010}* embryos at 72 h.p.f. **e**, *itga4* morphants could phenocopy *itga4^{cas005}*. The validated *itga4* morpholino oligonucleotide (MO) that can block the translation of *itga4* mRNA was injected into one-cell-stage wild-type embryos to produce *itga4* morphants. WISH results of *myb* expression in the control and *itga4* morphants at 72 h.p.f. **f**, **g**, Representative live

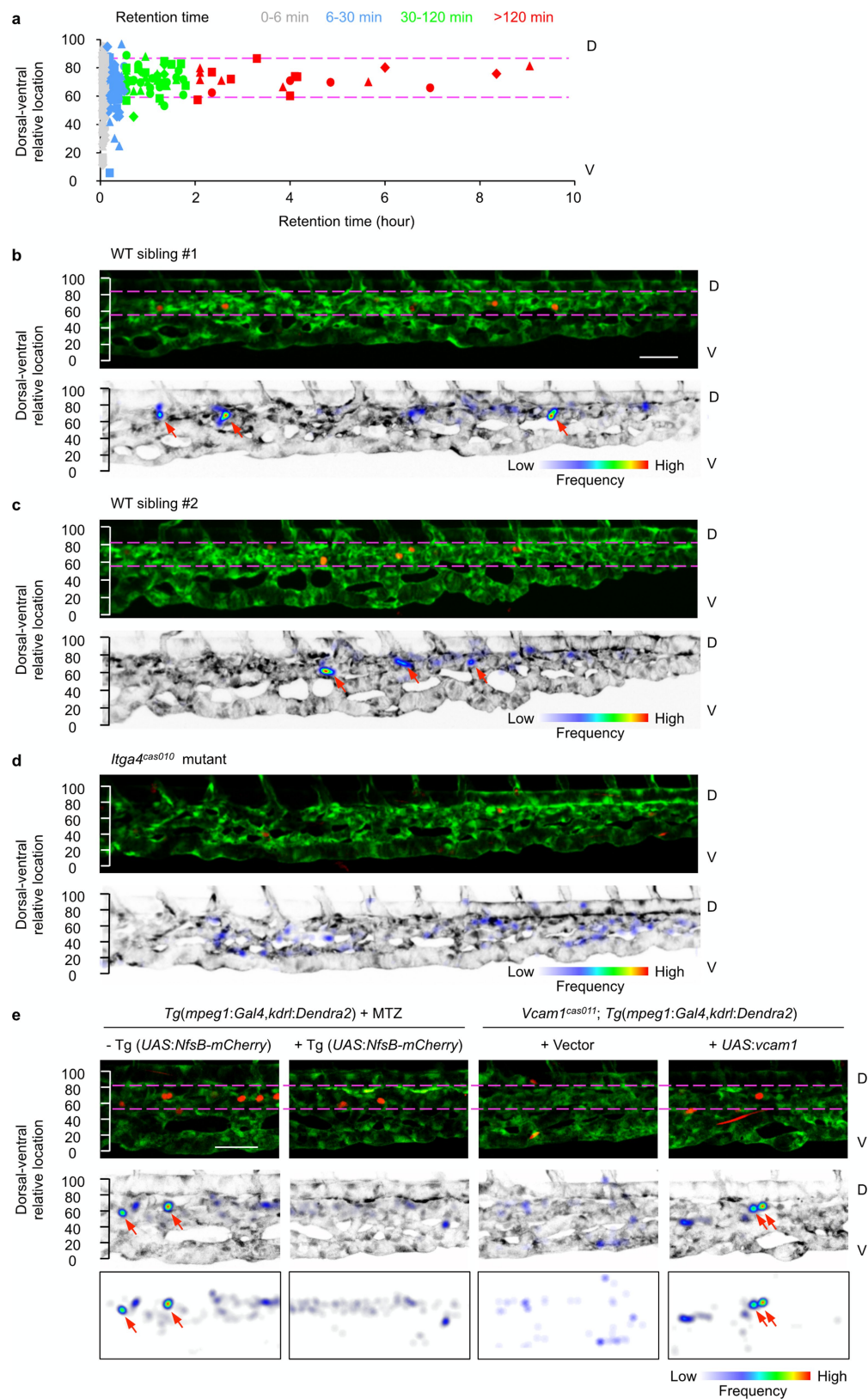
imaging (**g**) and statistical analysis (**f**) of *Tg(gata1:DsRed)*, *Tg(mpx:eGFP)* and *Tg(lyz:DsRed)* transgenic embryos after the injection of control or *itga4* or *vcam1* morpholino oligonucleotides at 84 h.p.f. Results show that downstream haematopoietic lineages were all defective owing to disrupted HSPC retention in the CHT. *gata1*, control vs *itga4* MO: **** $P < 0.0001$, $t = 7.42$, $df = 14$; *gata1*, control vs *vcam1* MO: **** $P < 0.0001$, $t = 8.11$, $df = 14$; *mpx*, control vs *itga4* MO: $P < 0.0001$, $t = 5.92$, $df = 14$; *mpx*, control vs *vcam1* MO: **** $P < 0.0001$, $t = 9.41$, $df = 14$; *lyz*, control vs *itga4* MO: **** $P < 0.0001$, $t = 7.16$, $df = 25$; *lyz*, control vs *vcam1* MO: **** $P < 0.0001$, $t = 5.74$, $df = 25$. **h**, In situ analysis of *itga4* expression in the AGM (FISH, 36 h.p.f.) and CHT (WISH, 72 h.p.f.) of *Tg(kdrl:eGFP)* (green) embryos after injection of control, *runx1* or *myb* morpholino oligonucleotides indicates HSPC cell-autonomous expression. **i–k**, *itga4* has an HSPC intrinsic mechanism during definitive haematopoiesis. **i**, The construction of the plasmid that was applied in the *itga4^{cas005}* mutant for Tol2-transposase-mediated transient transgenesis of *runx1*-enhancer-driven wild-type *itga4* expression. **j**, **k**, Phenotype analysis by *myb* WISH shows that the construct had a notable rescue effect on definitive haematopoiesis in the *itga4^{cas005}* mutant at 72 h.p.f. More than 45% of mCherry⁺ cells overlapped with *myb* FISH signalling in the CHT; a representative image is shown in **k**. Error bars denote s.e.m. Scale bars, 50 μm (**g**), 20 μm (**h**) and 5 μm (**k**).



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Photoconversion of *Tg(kdrl:Dendra2)* cells in the AGM can specifically mark HSPCs in the CHT. **a**, Schematic illustration (left) and confocal imaging analysis (right) of the HSPC labelling system in live *Tg(kdrl:Dendra2)* transgenic zebrafish larva. At 32–36 h.p.f., an ultraviolet laser was applied to photoconvert Dendra2 in haemogenic endothelium from green to red fluorescence in the area marked by the rectangle. Endothelial-to-haematopoietic transition was observed with egress of single red Dendra2⁺ cells (white arrows) from the aortic ventral wall into the sub-aortic space. **b**, Flow chart of the experimental analysis on HSPCs (red Dendra2⁺ cells) in live-imaging and confocal-imaging analysis after in situ RNA scope in identical locations. Green Dendra2 signalling could remain during the experiment, whereas red Dendra2 signalling could not (the red fluorescence of the ISV disappeared after in situ RNA scope analysis). Photoconverting the ISV in different embryos makes each embryo distinguishable. (We distinguished the embryos by the position of the photoconverted and weakened Dendra2 green ISV). After

live imaging the red Dendra2⁺ cells in the CHT (top, arrows) at 54 h.p.f., the embryos were fixed immediately for *runx1* in situ RNA scope analysis (bottom, yellow arrowheads). All HSPCs (red Dendra2⁺ cells) carry *runx1* transcripts, whereas there was no signal in the negative control. **c**, **d**, Confocal images of the CHT (**c**) and statistical analysis (**d**) at 54 h.p.f. show markedly reduced numbers of HSPCs (red Dendra2⁺ cells) in *runx1* and *myb* morphants, compared to that in control morphants. Control vs *runx1* MO: **** $P < 0.0001$, $t = 9.15$, $df = 9$; control vs *myb* MO: **** $P < 0.0001$, $t = 8.66$, $df = 9$. **e**, Live-imaging analysis on individual HSPCs in wild-type and *itga4^{cas005}* mutant embryos. Three representative frames from time-lapse imaging of 52–60 h.p.f. wild-type and *itga4^{cas005}* embryos with the *Tg(kdrl:Dendra2)* labelling system. The HSPCs (white arrow) remained stable in the vascular niche for over 30 min in the wild-type sibling (left); however, in the *itga4^{cas005}* mutants (right), the HSPCs (yellow arrowhead) remained for less than 9 min (see Supplementary Video 2). Error bars represent s.e.m. Scale bars, 50 μm (**a–c**) and 20 μm (**e**).

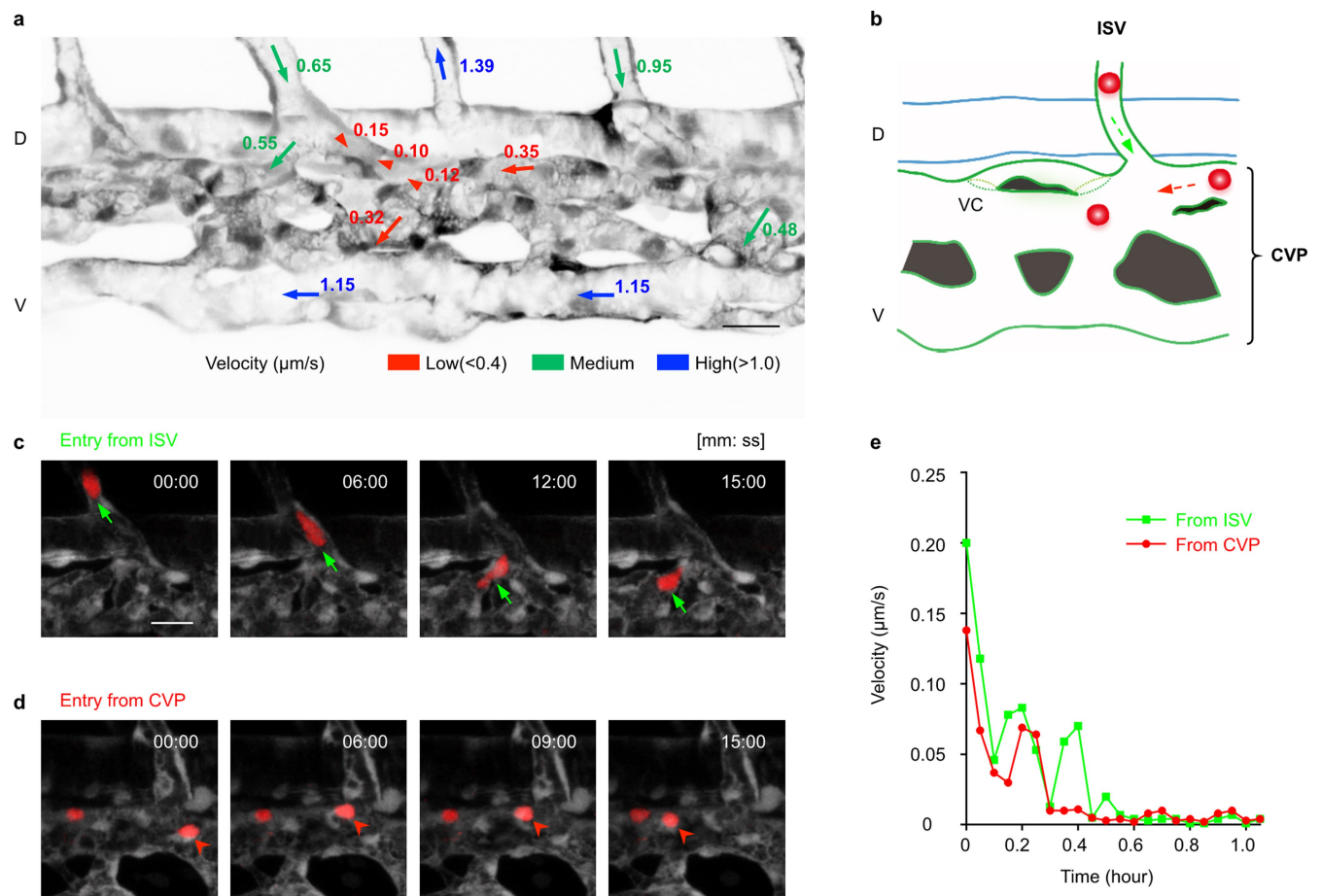


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | The HSPCs' retention 'hotspot' in the CHT.

a, Correlation analysis of retention time and the dorsal–ventral relative location of individual HSPCs in the CHT of four wild-type embryos at 50–60 h.p.f. Each shape represents one embryo (triangle, circle, square and rhombus), and each colour represent one class of retention time zone. HSPCs that remained for longer than 30 min were preferentially located in the region between the two magenta dashed lines enriched with venous capillaries. **b–d**, Longitudinal whole-mount images of the CHT in wild-type siblings (**b**, **c**) and *itga4^{cas010}* mutant embryos (**d**) in a *Tg(kdrl:Dendra2)* background after photoconversion and retention calculation of the frequency of HSPC appearance. **e**, Longitudinal whole-

mount images of the CHT in *Tg(mpeg1:Gal4,kdrl:Dendra2)* embryos after MTZ treatment with or without a *Tg(UAS:NfsB-mCherry)* background, and *vcam1^{cas011}*; *Tg(mpeg1:Gal4,kdrl:Dendra2)* embryos with transient transgenesis of vector (*UAS:polyA*) or *UAS:vcam1* after photoconversion and retention calculation of the frequency of HSPC appearance. Red arrows denote retention hotspots. HSPCs were preferentially located in the region between the two magenta dashed lines. Retention hotspots disappeared in the *itga4^{cas010}* and *vcam1^{cas011}* mutants after depletion of macrophages (MTZ treatment on NfsB-expressing macrophages). Scale bars, 50 μ m.

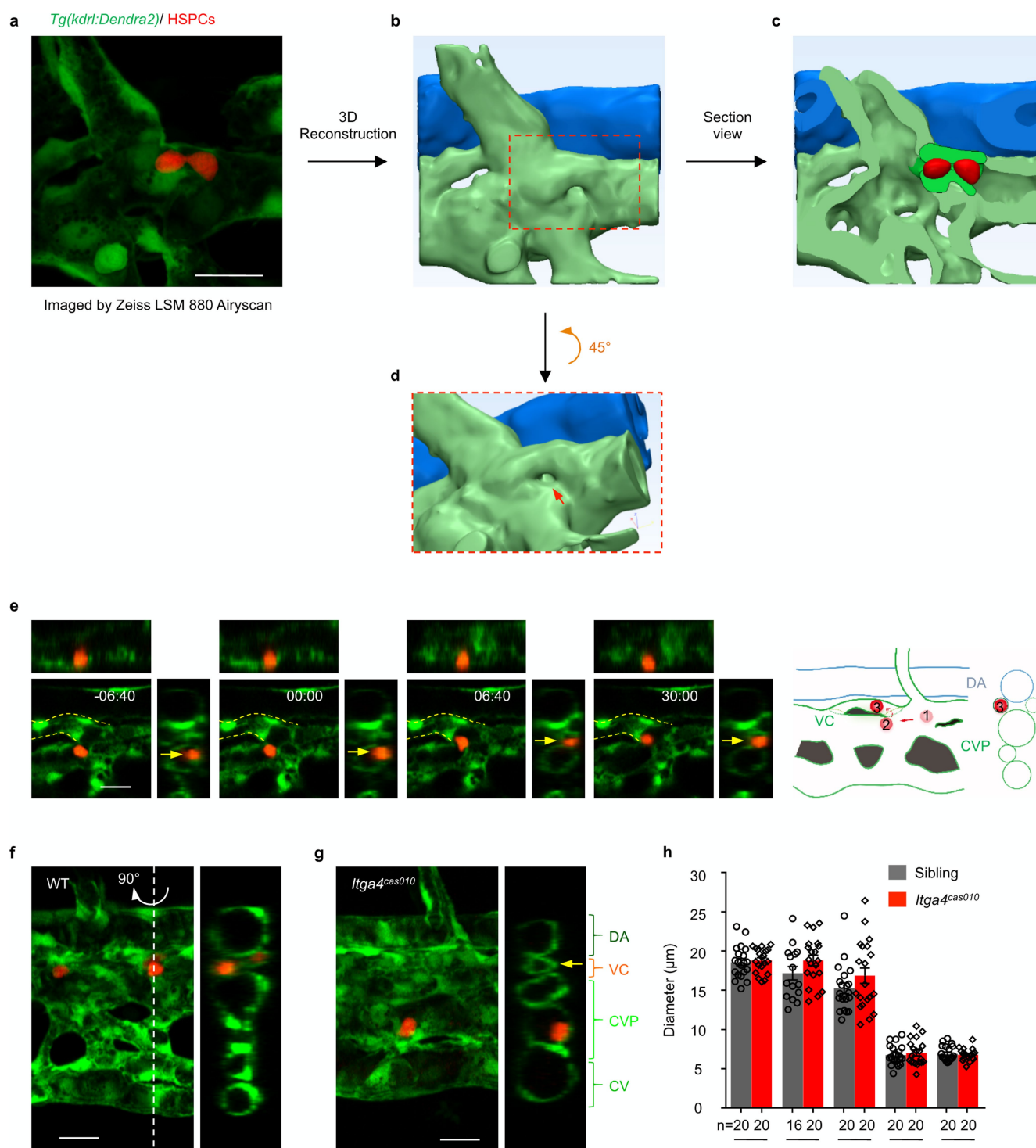


Extended Data Fig. 5 | HSPCs decelerate in the CHT vascular niche.

a, Representative spatial maps of the flow velocity of photoconverted HSPCs in the caudal vasculature at 54 h.p.f. Arrows show the direction of blood flow, and different colours indicate the level of velocity.

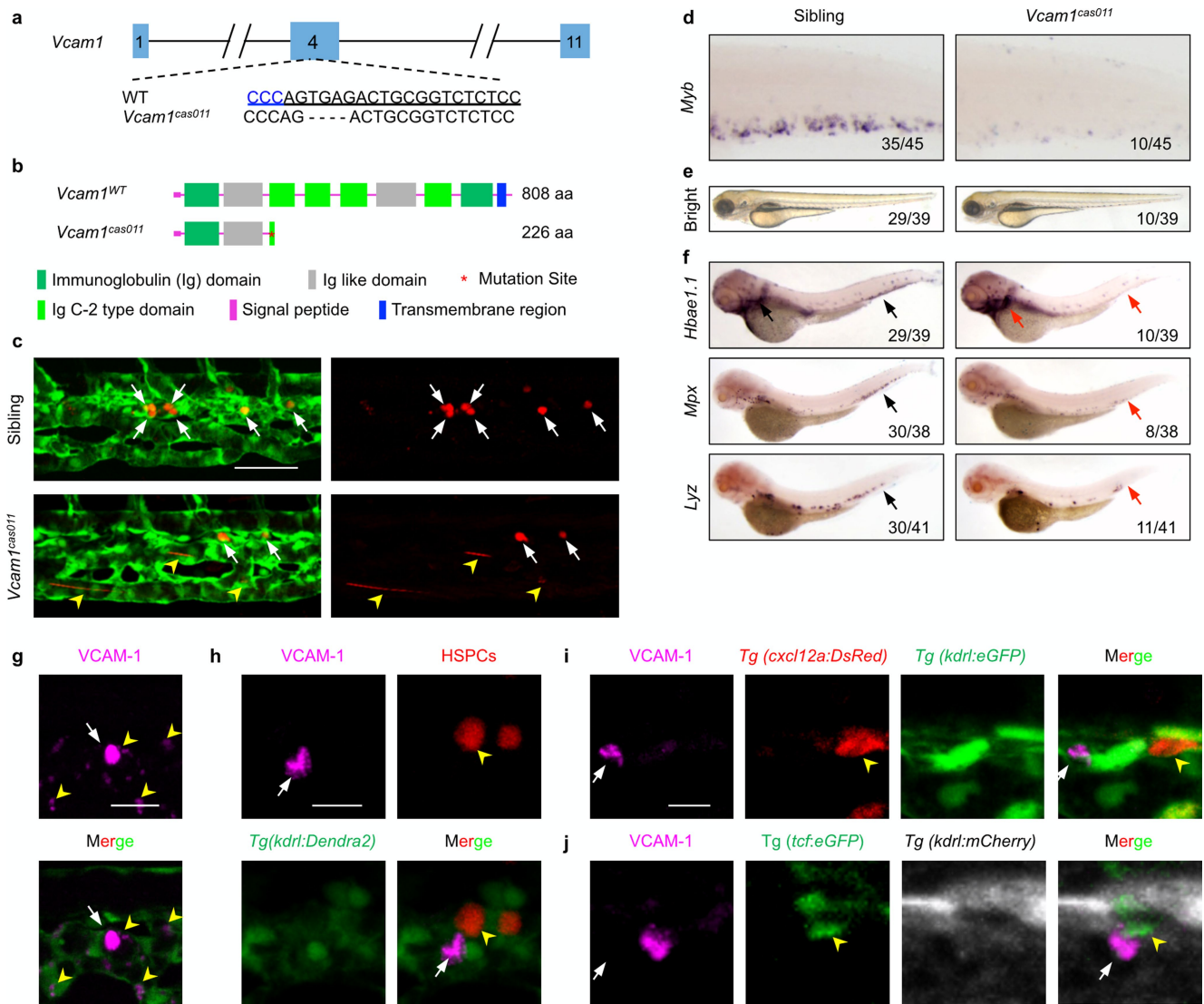
b, Schematic illustration of how HSPCs enter the CHT via the circulation.

c–e, Time-lapse imaging (**c**, **d**) and speed curve diagram (**e**) show that HSPCs arrive either from the intersegmental vessel (green arrow in **c**) or the CVP (red arrowhead in **d**) into the CHT and gradually decelerate (see Supplementary Video 3). Scale bar, 20 μm .



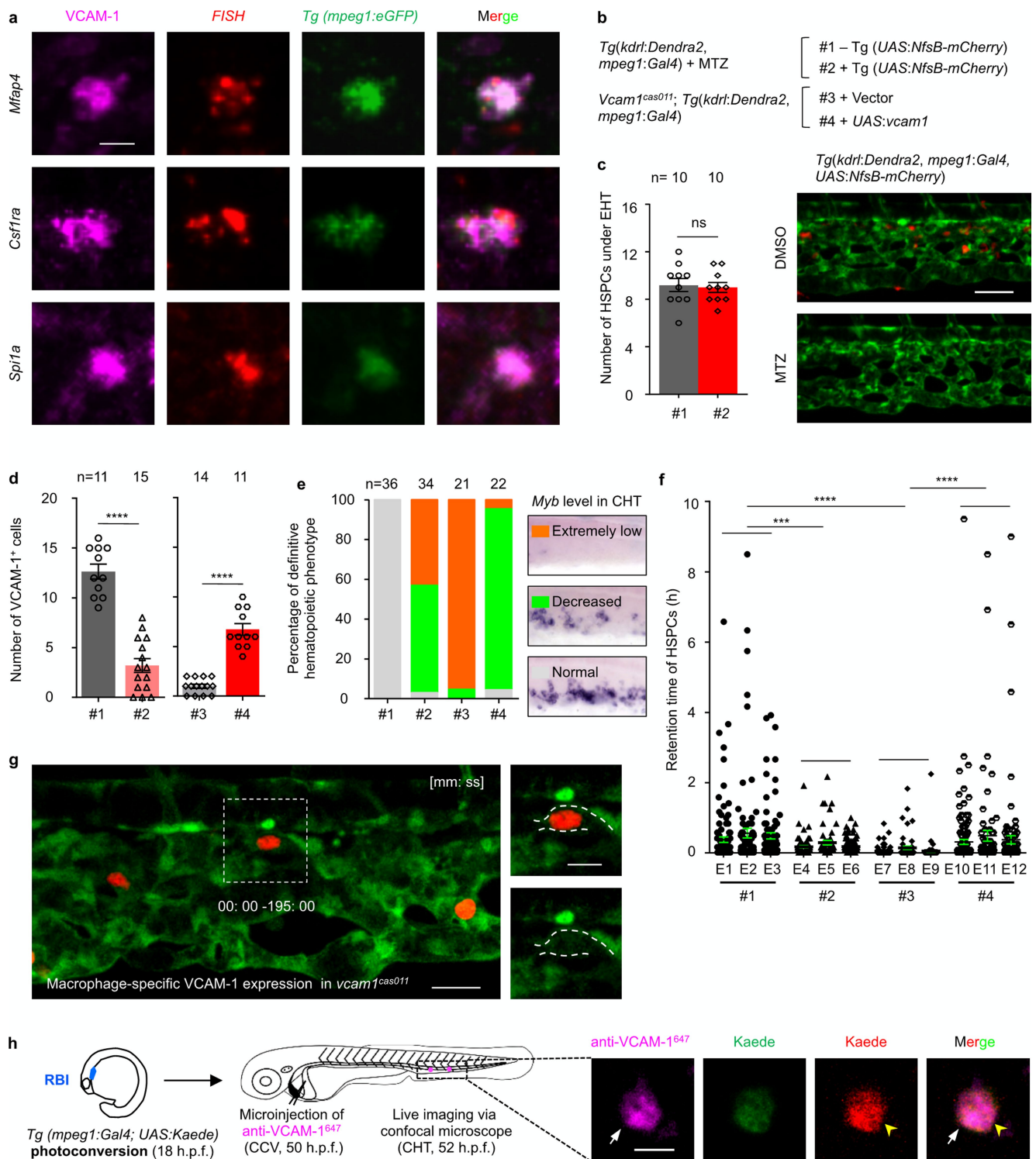
Extended Data Fig. 6 | The representative high-resolution vascular structure and HSPC in the retention hotspot. a, The original fluorescent image of the vessel surrounding HSPCs in the retention hotspot was captured by an LSM880 microscope equipped with Airyscan function and processed by 3D reconstruction (see Supplementary Video 4). **b**, 3D reconstruction of **a**. **c**, Section view of the caudal vein plexus and capillary. **d**, The 45° rotation view of the red frame in **b**. **e**, Time-lapse imaging (left) and scheme graph (right) show how HSPC retention occurs. HSPCs

initially came into the venous plexus and then entered the venous capillary for long-term retention. **f–h**, Images (**f**, **g**) and statistical analysis (**h**) show that the diameter of various vessels in the CHT in *itga4^{cas010}* mutants (**g**) is similar to that in the wild-type siblings (**f**) at 54 h.p.f. The inner diameter of venous capillaries, but not other vessels, is close to the diameter of HSPCs. DA: $P = 0.73$, $t = 0.35$, $df = 19$; CV: $P = 0.14$, $t = 1.52$, $df = 33$; CVP: $P = 0.17$, $t = 1.41$, $df = 19$; VC: $P = 0.63$, $t = 0.49$, $df = 19$; HSPC: $P = 0.67$, $t = 0.44$, $df = 19$. Scale bar, 20 μm.



Extended Data Fig. 7 | Characterization of VCAM-1⁺ cells in the CHT.
a, Generation of the *vcam1* mutant using the CRISPR-Cas9 technique. The alignment of wild-type (underlined) and mutated sequences is listed. The PAM sequence of gRNA is 'GGG' (in blue). Deletions are indicated by dashes. **b**, According to the stop codon in the genome, SMART software was used to predict the structure of the wild-type *vcam1* and *vcam1^{cas011}* presumed protein. The molecular sizes of the presumed protein are indicated. **c**, Live imaging of the CHT at 54 h.p.f. shows retention defects in *vcam1^{cas011}* mutants. Representative images show that most HSPCs resided within the CHT (white arrows) in wild-type siblings (top), whereas these cells went through quickly in *vcam1^{cas011}* embryos (bottom, yellow arrowheads) (see Supplementary Video 5). **d**, WISH analysis of *myb* expression in the CHT of wild-type and *vcam1^{cas011}* mutant zebrafish embryos. **e, f**, The definitive haematopoiesis is defective in *vcam1^{cas011}*

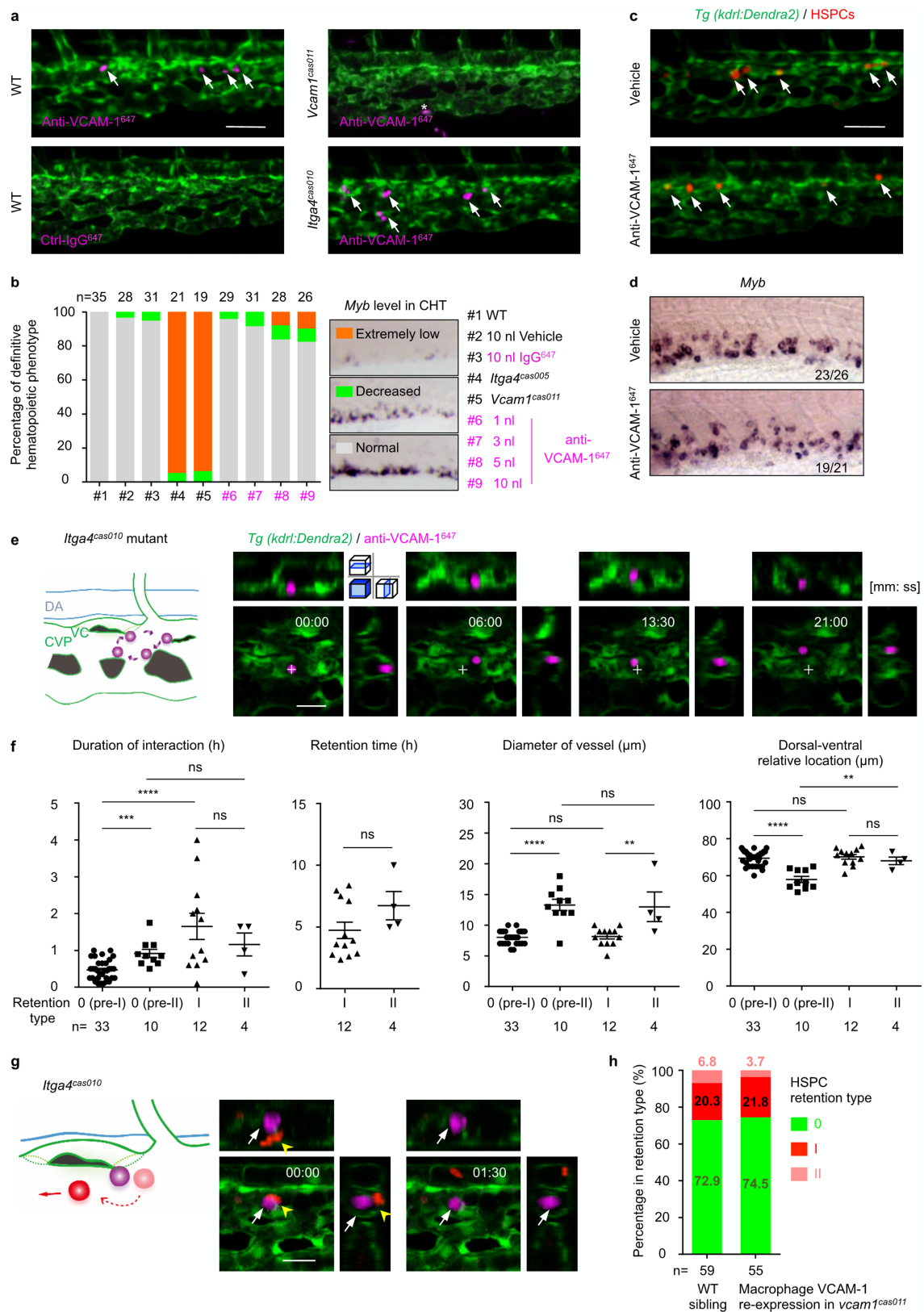
mutant zebrafish embryos. **e**, The bright-field images of wild-type and *vcam1^{cas011}* embryos show no obvious morphological difference at 72 h.p.f. **f**, WISH results of *hbae1.1*, *mpx* and *lyz* expression in wild-type and *vcam1^{cas011}* mutant embryos at 72 h.p.f. Arrows indicate the comparable position in wild-type (black) or *vcam1^{cas011}* (red) embryos. **g**, Magnified views showed VCAM-1 was mainly expressed in individual cells (white arrow) but weakly expressed on the venous endothelial cells (yellow arrowheads). **h**, After photoconversion, *Tg(kdrl:Dendra2)* embryos are stained with anti-VCAM-1 (magenta, white arrow). The yellow arrowhead denotes an HSPC. **i**, *Tg(cxcl12a:DsRed,kdrl:eGFP)* transgenic embryos are stained with anti-VCAM-1 (magenta, white arrow) and anti-DsRed (red, yellow arrowhead). **j**, *Tg(tcf:eGFP,kdrl:mCherry)* transgenic embryos are stained with anti-VCAM-1 (magenta, white arrows) and anti-GFP (green, yellow arrowheads). Scale bars, 50 μ m (c), 20 μ m (g) and 10 μ m (h, i).



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Distinct role of macrophages and venous endothelium VCAM-1 in HSPCs retention. **a**, Representative FISH confocal imaging of *mfap4* (top), *csf1ra* (middle), *spi1a* (bottom) immunofluorescence with anti-VCAM-1 and anti-GFP antibodies indicates that VCAM-1⁺ cells in the CHT are macrophage-like cells (see Supplementary Table 2). **b**, The construction of the plasmid applied in **c–f**. **c**, Validation of the macrophage-specific cell-depletion system. Left, the number of HSPCs under the endothelial-to-haematopoietic transition (EHT) process in the AGM of *Tg(mpeg1:Gal4,kdrl:Dendra2)* transgenic embryos at 54 h.p.f. with MTZ treatment and with (#2) or without (#1) the *Tg(UAS:NfsB-mCherry)* background. $P = 0.80$, $t = 0.25$, $df = 9$. Right, live imaging of vessels (green) and macrophages (red) with or without MTZ treatment in the CHT of *Tg(kdrl:Dendra2,mpeg1:Gal4,UAS:NfsB-mCherry)* transgenic embryos showed that MTZ treatment could delete almost all mCherry⁺ macrophages. **d**, Quantification of VCAM-1⁺ cells in the CHT, detected by anti-VCAM-1 immunofluorescence, in *Tg(mpeg1:Gal4,kdrl:Dendra2)* embryos at 54 h.p.f. with MTZ treatment and with (#2) or without (#1) a *Tg(UAS:NfsB-mCherry)* background, and in *vcam1^{cas011}* mutants with Tol2-mediated transient transgenesis of vector (*UAS:polyA*) (#3) or *UAS:vcam1* (#4) in a *Tg(mpeg1:Gal4,kdrl:Dendra2)* background. #1 vs #2: **** $P < 0.0001$, $t = 9.18$, $df = 24$; #3 vs #4: **** $P < 0.0001$, $t = 10.03$, $df = 23$. **e**, Statistical analysis shows the percentage of the three types of definitive haematopoietic phenotype

(extremely low, decreased or normal) in the four experimental conditions (#1–#4) linked to **d**. Macrophage-specific cell depletion caused deficient definitive haematopoiesis; however, macrophage-specific VCAM-1 re-expression markedly rescued deficient haematopoiesis in *vcam1^{cas011}* mutants. **f**, Retention time of individual HSPCs in transgenic *Tg(mpeg1:Gal4,kdrl:Dendra2)* embryos at 50–60 h.p.f. with MTZ treatment and with (#2) or without (#1) a *Tg(UAS:NfsB-mCherry)* background, and in *vcam1^{cas011}* mutants with Tol2-mediated transient transgenesis of vector (*UAS:polyA*) (#3) or *UAS:vcam1* (#4) in a *Tg(mpeg1:Gal4,kdrl:Dendra2)* background (see Fig. 2f, g). #1 vs #2: *** $P = 0.0001$, $t = 3.85$, $df = 550$; #1 vs #3: **** $P < 0.0001$, $t = 6.05$, $df = 565$; #3 vs #4: **** $P < 0.0001$, $t = 4.37$, $df = 590$. **g**, Live-imaging frame shots of HSPCs in which macrophage-specific VCAM-1 was in re-expressed *vcam1^{cas011}* mutants from Fig. 2g (see Supplementary Video 6). **h**, Schematic illustration shows that macrophage labelling (photoconverted Kaede⁺; red) was performed at 18 h.p.f. in the rostral blood island (RBI) in *Tg(mpeg1:Gal4,UAS:Kaede)* embryos, followed by a 1 nl anti-VCAM-1⁶⁴⁷ (0.4 ng) antibody injection at 50 h.p.f. Cell-lineage tracing of the labelled macrophages (red; yellow arrowheads) in vivo was performed from 2 h after the injection. Representative images show that macrophages from the rostral blood island at 18 h.p.f. migrate to the CHT, and are VCAM-1⁺. Scale bars, 50 μm (**c**), 20 μm (**g**), 10 μm (**h**) and 5 μm (**a**).



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Anti-VCAM-1⁶⁴⁷ antibody labels usher cells without disrupting definitive haematopoiesis. **a**, Injection of 1 nl (0.4 ng) of anti-VCAM-1⁶⁴⁷ antibody labels usher cells (arrows) in wild-type and *itga4^{cas010}* mutants in the *Tg(kdr:eGFP)* background, whereas injection of either control (non-specific) IgG⁶⁴⁷ antibody into wild-type cells or anti-VCAM-1⁶⁴⁷ antibody into *vcam1^{cas011}* mutants in the *Tg(kdr:eGFP)* background did not label any cells in the retention hotspots. Asterisk indicates nonspecific labelling on a chromatophore in the CHT. **b**, Anti-VCAM-1⁶⁴⁷ injection marginally influence definitive haematopoiesis. Statistical analysis shows the percentage of the three types of definitive haematopoietic phenotype in nine different conditions, including wild-type embryos without injection (#1), with 10 nl vehicle (#2) or 10 nl 0.4 mg ml⁻¹ IgG⁶⁴⁷ injection (#3), *itga4^{cas005}* mutants (#4) or *vcam1^{cas011}* mutants (#5) without injection, and wild-type embryos with 1–10 nl 0.4 mg ml⁻¹ anti-VCAM-1⁶⁴⁷ injection (#6–#9). **c**, **d**, Live imaging of HSPCs (**c**) or WISH analysis of the *myb* probe at 60 h.p.f. (**d**) of the wild-type CHT after vehicle or 1 nl anti-VCAM-1⁶⁴⁷ antibody injection. **e**, Schematic diagrams (left) and confocal imaging (right) show VCAM-1⁺ cells patrolling on a small scale in the CHT in *itga4^{cas010}* mutant embryos. Cross indicates the original position at the initial time point. **f**, Statistical analysis of the duration of the interaction between HSPCs and usher cells,

the HSPC retention time, the diameter of vessels and the dorsal–ventral relative location for HSPC retention in pre-type I (type 0), pre-type II (type 0), type I and type II. Duration, pre-I vs pre-II: *** $P = 0.0001$, $t = 4.25$, $df = 41$; duration, pre-I vs I: **** $P < 0.0001$, $t = 5.31$, $df = 43$; duration, pre-II vs II: $P = 0.37$, $t = 0.93$, $df = 12$; duration, I vs II: $P = 0.46$, $t = 0.76$, $df = 14$; retention: $P = 0.16$, $t = 1.50$, $df = 14$; diameter, pre-I vs pre-II: **** $P < 0.0001$, $t = 8.80$, $df = 41$; diameter, pre-I vs I: $P = 0.68$, $t = 0.42$, $df = 43$; diameter, pre-II vs II: $P = 0.89$, $t = 0.14$, $df = 12$; diameter, I vs II: ** $P = 0.006$, $t = 3.24$, $df = 14$; location, pre-I vs pre-II: **** $P < 0.0001$, $t = 7.64$, $df = 41$; location, pre-I vs I: $P = 0.6$, $t = 0.53$, $df = 43$; location, pre-II vs II: ** $P = 0.005$, $t = 3.41$, $df = 12$; location, I vs II: $P = 0.41$, $t = 0.85$, $df = 14$. **g**, In *itga4^{cas010}* mutants, HSPCs encountered but failed to interact with usher cells and then went through the CHT within a few minutes (see Supplementary Video 10). **h**, The percentage of the type 0, type I and type II HSPC retention types in wild-type sibling and *vcam1^{cas011}* mutants in the *Tg(mpeg1:Gal4,kdr:Dendra2)* background with transient transgenesis of *UAS:vcam1*. None of the HSPCs in the *vcam1* mutants could be classified into either type I or II retention types, or were comparable with the HSPCs in Extended Data Fig. 8h. Scale bars, 50 μ m (**a**, **c**) and 20 μ m (**e**, **g**).

TIC236 links the outer and inner membrane translocons of the chloroplast

Yih-Lin Chen¹, Lih-Jen Chen¹, Chiung-Chih Chu¹, Po-Kai Huang^{1,2}, Jie-Ru Wen¹ & Hsoun-min Li^{1*}

The two-membrane envelope is a defining feature of chloroplasts. Chloroplasts evolved from a Gram-negative cyanobacterial endosymbiont. During evolution, genes of the endosymbiont have been transferred to the host nuclear genome. Most chloroplast proteins are synthesized in the cytosol as higher-molecular-mass preproteins with an N-terminal transit peptide. Preproteins are transported into chloroplasts by the TOC and TIC (translocons at the outer- and inner-envelope membranes of chloroplasts, respectively) machineries^{1,2}, but how TOC and TIC are assembled together is unknown. Here we report the identification of the TIC component TIC236; TIC236 is an integral inner-membrane protein that projects a 230-kDa domain into the intermembrane space, which binds directly to the outer-membrane channel TOC75. The knockout mutation of *TIC236* is embryonically lethal. In *TIC236*-knockdown mutants, a smaller amount of the inner-membrane channel TIC20 was associated with TOC75; the amount of TOC–TIC supercomplexes was also reduced. This resulted in a reduced import rate into the stroma, though outer-membrane protein insertion was unaffected. The size and the essential nature of TIC236 indicate that—unlike in mitochondria, in which the outer- and inner-membrane translocons exist as separate complexes and a supercomplex is only transiently assembled during preprotein translocation^{3,4}—a long and stable protein bridge in the intermembrane space is required for protein translocation into chloroplasts. Furthermore, TIC236 and TOC75 are homologues of bacterial inner-membrane TamB⁵ and outer-membrane BamA, respectively. Our evolutionary analyses show that, similar to TOC75, TIC236 is preserved only in plants and has co-evolved with TOC75 throughout the plant lineage. This suggests that the backbone of the chloroplast protein-import machinery evolved from the bacterial TamB–BamA protein-secretion system.

To identify novel proteins associated with TOC75, we used solubilized membranes of isolated root leucoplasts⁶—rather than leaf chloroplasts—for co-immunoprecipitation (Extended Data Fig. 1) to avoid extremely abundant photosynthetic proteins hindering protein detection. Among the proteins we identified was the pea homologue of *Arabidopsis* EMB2410 (At2g25660). Knockout mutation of *EMB2410* was embryonically lethal⁷. *EMB2410* is expressed in all tissues and is predicted to encode a protein of unknown function of 2,166 amino acids with a plastid-targeting transit peptide, an N-terminal transmembrane domain and a C-terminal ‘domain of unknown function 490’ (DUF490) (Fig. 1a and Extended Data Fig. 2). EMB2410 homologues are present in all plants (see below) and may be homologues of TamB⁵, which is found in Gram-negative bacteria. A missense mutation in the rice orthologue *SSG4* results in defective endosperm amyloplast starch grains⁸. In Gram-negative bacteria, TamB is anchored in the inner membrane and protrudes across the periplasmic space to directly interact with the outer-membrane protein TamA or BamA. TOC75, TamA and BamA are all members of the Omp85 family. They are composed of a C-terminal β -barrel domain that spans the membrane, which is preceded by various numbers of polypeptide transport-associated (POTRA; see Supplementary Discussion) domains that extend into

the intermembrane or periplasmic space^{9,10}. Several autotransporter proteins—for example, adhesins—are secreted by the bacterial TamB–TamA or TamB–BamA system^{5,11}.

Antibodies prepared against EMB2410 recognized an inner-envelope membrane protein of about 240 kDa in size (Fig. 1b). EMB2410 is easily degraded to a protein of about 220 kDa in size (blue dot in Fig. 1b, and in all figures thereafter), so it often appears as a doublet on gels. EMB2410 is resistant to alkaline extraction of chloroplast membranes (Fig. 1c), and resistant to thermolysin but sensitive to trypsin digestion in intact chloroplasts (Fig. 1d). Because our antibodies recognize the C-terminal portion of EMB2410, these results indicate that EMB2410 is anchored in the inner-envelope membrane—oriented with its N-terminal in the stroma and C-terminal in the intermembrane space—and that the great majority of the polypeptide of this protein is located in the intermembrane space. To verify whether EMB2410 possesses a cleavable transit peptide and to better establish the size of the mature protein, we performed in vitro protein-import experiments using a C-terminal-truncated construct (Extended Data Fig. 3). The results show that EMB2410 has a transit peptide of approximately 37 residues in length; the calculated molecular mass of EMB2410 after transit peptide removal is 236 kDa. Hereafter, we name EMB2410 as TIC236.

Next, we investigated whether TIC236 is associated with other translocon components. Chloroplast membranes were solubilized and immunoprecipitated with anti-TOC75 antibodies. TIC236 was detected in the immunoprecipitates together with TIC110, TIC40 and TIC20 but not with the inner-envelope membrane protein IEP37, which is not part of the translocon (Fig. 1e). The results were confirmed by performing co-immunoprecipitations with anti-TIC110 or anti-TIC236 antibodies (Fig. 1e). We then solubilized and fractionated chloroplasts that contained translocating prRBCS—the preprotein of the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO)—radiolabelled with ³⁵S ([³⁵S]prRBCS), using sucrose density gradients. The result shows that TIC236 has a similar sedimentation pattern to that of TOC75: a portion of total TIC236—that included the degraded form—was detected at the top of the gradient and about half was detected in fractions 14 to 18, which also contained TOC159, TOC75, TIC110, TIC20 and translocating [³⁵S]prRBCS (Fig. 1f and Extended Data Fig. 4a–c). Two stable TOC–TIC supercomplexes have previously been observed by blue native polyacrylamide gel electrophoresis (BN-PAGE), in 1- and 1.25-megadalton forms that represent the major forms in pea and *Arabidopsis* chloroplasts, respectively¹². We therefore performed two-dimensional BN-PAGE analyses of pea chloroplasts. A portion of TIC236 was detected in the 1-megadalton TOC–TIC supercomplex region together with TOC75 and TIC20, and the remaining TIC236 (including all the degraded TIC236) was located in a region below 440 kDa (Fig. 1g and Extended Data Fig. 4d). These data support the conclusion that a substantial proportion of TIC236 is associated with other translocon components in the TOC–TIC supercomplexes.

To investigate the function of TIC236, we identified three *Arabidopsis* mutants that have knockout or knockdown mutations in *TIC236*. The *tic236-1* allele (SALK_048770, *Columbia* (*Col*) ecotype) contains

¹Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan. ²Present address: Department of Plant Sciences, University of California, Davis, CA, USA. *e-mail: mbhmli@gate.sinica.edu.tw

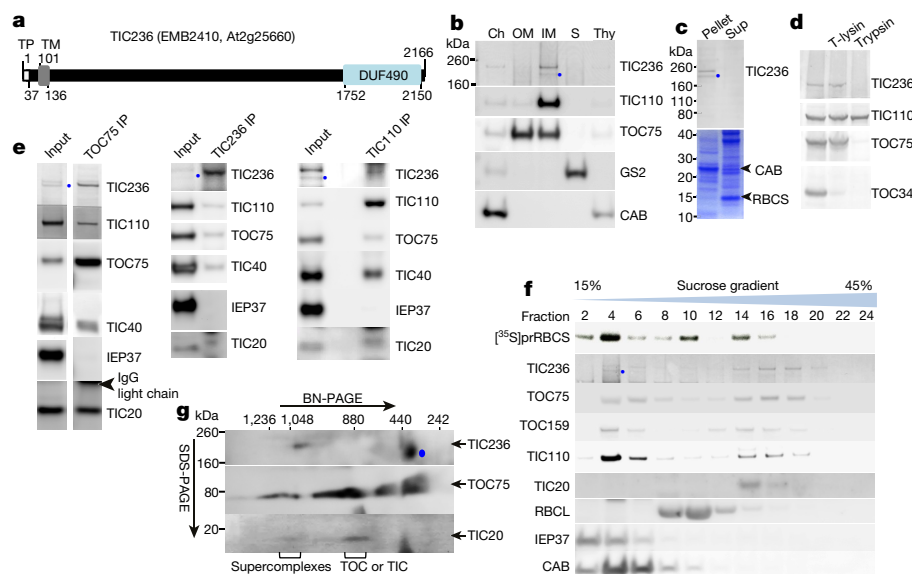


Fig. 1 | TIC236 is an integral inner-membrane protein and a member of the TOC-TIC supercomplexes. All panels show results of SDS-PAGE and immunoblotting with antibodies as labelled, except where indicated. **a**, Schematic of domain structure of *Arabidopsis* TIC236. TP (white box), transit peptide. TM (grey box), predicted transmembrane domain. The numbers designate the amino acid residue number, with the initiation methionine of TIC236 preprotein as 1. **b**, Pea chloroplasts (Ch) were fractionated into the outer (OM) and inner (IM) envelope membranes, stroma (S) and thylakoids (Thy). The blue dot indicates a degradation product of TIC236. GS2, stromal glutamine synthetase. CAB, thylakoid chlorophyll *a/b* binding protein. **c**, Pea chloroplasts were separated into pellet and supernatant (sup) fractions by alkaline extraction. Bottom

panel shows samples stained with Coomassie blue. **d**, Pea chloroplasts (left lane) were treated with thermolysin (T-lysin) or trypsin. **e**, Membranes of chloroplasts treated with 1 mM dithiobis(succinimidyl propionate) (DSP) were solubilized with 1% decylmaltoside (input) and immunoprecipitated (IP) with anti-TOC75, anti-TIC110 or anti-TIC236 antibodies. **f**, Pea chloroplasts that contain translocating [35 S]prRBCS were solubilized by 1% decylmaltoside and analysed by a 15–45% linear sucrose density gradient. Fractions were analysed by SDS-PAGE followed by fluorography for [35 S]prRBCS. RBCL, the large subunit of RuBisCO. **g**, Pea chloroplasts were solubilized with 1% digitonin and analysed by 2D BN-PAGE. Data shown are representative of at least two independent experiments (**b–g**).

a transfer DNA (T-DNA) insertion at the tenth exon (Fig. 2a). No homozygous mutant of this allele could be obtained. Approximately a quarter of the seeds of heterozygous plants are albino or shrunken (Fig. 2b), which confirms that a knockout mutation in *TIC236* causes embryonic lethality⁷. The *tic236-2* (SAIL104-F07, *Col* ecotype) and *tic236-3* (RIKEN PST00216, *Nossen* (*No-0*) ecotype) alleles have a T-DNA and a Ds transposon insertion in the 5' untranslated region (UTR), respectively, each of which results in severely reduced levels of *TIC236* RNA and *TIC236* protein (Extended Data Fig. 5). Both of these 5'-UTR mutants are much smaller than their corresponding wild types (Fig. 2c). Their leaves have lesions and irregular margins that are most probably caused by cell death during leaf development.

To examine whether *TIC236* functions in protein import, four preproteins that are destined for different locations within chloroplasts—prTIC40 (inner-envelope membrane), prRBCS and prHSP93 (stroma) and prOE23 (thylakoid lumen)—were imported into chloroplasts isolated from the *tic236-2* and *tic236-3* mutants, and their corresponding wild types. Both mutants exhibited reduced import of all four preproteins (Fig. 2d). Mutant chloroplasts often showed increased preprotein accumulation. These preproteins could be degraded by thermolysin (Extended Data Fig. 6a), which indicates that they were still exposed on the chloroplast surface. Import time-course experiments revealed that the mutant chloroplasts had reduced import rates (Fig. 2f and Extended Data Fig. 6b). The amounts of major translocon components in the *tic236* mutant chloroplasts were then measured by immunoblots, and shown not to be reduced (Extended Data Fig. 6c); this confirms that the reduced import observed was not due to a secondary effect of lower amounts of other translocon components. We next examined the insertion of two outer-membrane proteins, TOC34 from *Arabidopsis* and OEP14 from pea. OEP14 insertion relies solely on TOC75, without the need for any inner-membrane proteins¹³. After insertion, both TOC34 and OEP14 produced signature fragments—of 6 kDa and 3 kDa in size, respectively^{13,14}—that were protected by the outer membrane after thermolysin digestion (Fig. 2e, arrowheads). As shown in Fig. 2e, the

insertion of TOC34 and OEP14 was not reduced in the *tic236* mutant chloroplasts.

The DUF490 domain of *Escherichia coli* TamB interacts directly with the POTRA domains of TamA; deletion of seven residues from the C terminus of TamB DUF490 compromises this interaction⁵. We tested whether the DUF490 domain of *TIC236* also interacts with the POTRA domains of TOC75. The DUF490 domain of *TIC236*—or the domain minus the last 16 residues—was fused to the C terminus of glutathione S-transferase (GST) to create GST-DUF490 or GST-DUF490 Δ C, respectively. A His₆ tag was fused to the C terminus of the three POTRA domains of *Arabidopsis* TOC75 to create POTRA1-POTRA2-POTRA3-His₆⁹. GST-DUF490, but not GST, was pulled down by POTRA1-POTRA2-POTRA3-His₆ (Fig. 3a). Deleting the last 16 residues from the C-terminal of DUF490 reduced the association by 50%. Constructs that consisted of only POTRA1-POTRA2 or POTRA1 were then made. POTRA1-POTRA2 bound DUF490 equally as well as did POTRA1-POTRA2-POTRA3, whereas POTRA1 alone exhibited limited binding (Extended Data Fig. 7a); this suggests that POTRA3 is dispensable, and POTRA2 important, for interacting with DUF490.

We next tested whether *TIC236* directly interacts with TOC75 *in vivo*. Isolated *Arabidopsis* chloroplasts were treated with the heterobifunctional crosslinker succinimidyl 4-(*N*-maleimidomethyl) cyclohexane-1-carboxylate (SMCC, spacer arm 8.3 Å), solubilized with 1% lithium dodecyl sulfate (LDS) to disrupt non-covalent protein-protein interactions, and immunoprecipitated with anti-TOC75 antibodies. The immunoprecipitates were hybridized with anti-TOC75 or anti-TIC236 antibodies. In the anti-TOC75 immunoprecipitates, anti-TIC236 antibody recognized a protein complex of approximately 300 kDa in size (Fig. 3b, lane 4, and Extended Data Fig. 7b), suggesting that it is an adduct of TOC75 plus *TIC236*. Because *TIC236* is part of the TOC-TIC supercomplexes (Fig. 1), the crosslinking of *TIC236* with TOC75 indicates that *TIC236* provides a physical link between the TOC and TIC translocons. We further used immunogold electron

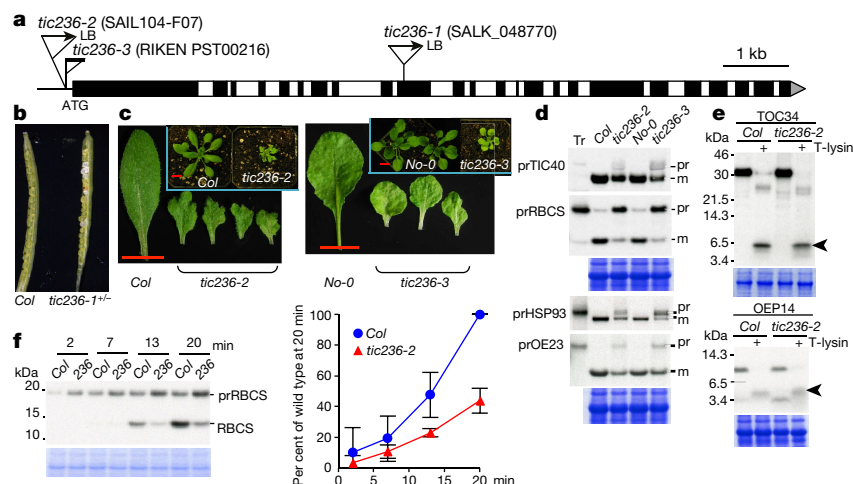


Fig. 2 | *tic236*-mutant chloroplasts exhibit reduced import into the stroma but not into the outer membrane. **a**, Exon and intron organization of *TIC236* and the T-DNA or transposon insertion positions of the three *tic236*-mutant alleles. Black and white boxes represent exons and introns, respectively. Grey triangle, 3' UTR. The exact length of the *TIC236* 5' UTR has not been determined. LB, left border of T-DNA; ATG, the translation initiation methionine. **b**, Siliques from *tic236-1* heterozygous (*tic236-1^{+/-}*) and wild-type (*Col*) plants. **c**, Leaf phenotypes of *tic236-2* and *tic236-3* mutants and their corresponding wild types (*Col* and *No-0*, respectively). Inserts, 24-day-old seedlings. Scale bar (red lines), 1 cm. **d**, Preproteins (Tr) were imported into chloroplasts

isolated from *tic236-2* or *tic236-3*, and the corresponding wild types (*Col* or *No-0*, respectively). The region of the gel around CAB is shown below the fluorograph. prTIC40 and prRBCS were imported together into the same chloroplast preparation, as were prHSP93 and prOEP23. pr, precursor form of the respective preprotein; m, mature form of the respective preprotein. **e**, TOC34 and OEP14 were imported into isolated wild-type (*Col*) and *tic236-2* chloroplasts and treated with thermolysin (T-lysin). **f**, [³⁵S]prRBCS was imported into wild-type (*Col*) and *tic236-2* (236) chloroplasts for various amounts of time. The amount of imported RBCS was quantified. Data shown as mean \pm s.d. of three independent experiments (**f**), and representative of three (**e**) and two (**b–d**) independent experiments.

microscopy to confirm the localization of TIC236. Gold particles were primarily found along the periphery of the chloroplast envelope and could be observed at contact sites between the two envelope membranes (Fig. 3c, d).

If TIC236 serves as the link between the TOC and TIC complexes, then the association between TOC and TIC should be reduced in *tic236*

knockdown mutants. Indeed, much less TIC20 was co-immunoprecipitated with TOC75 in the *tic236*-mutant chloroplasts (Fig. 3e). We next used BN-PAGE to analyse the amount of [³⁵S]prRBCS that was associated with the supercomplexes during import. For all ATP concentrations we assayed, the amount of prRBCS that was associated with the supercomplexes was much lower in the mutant than in the

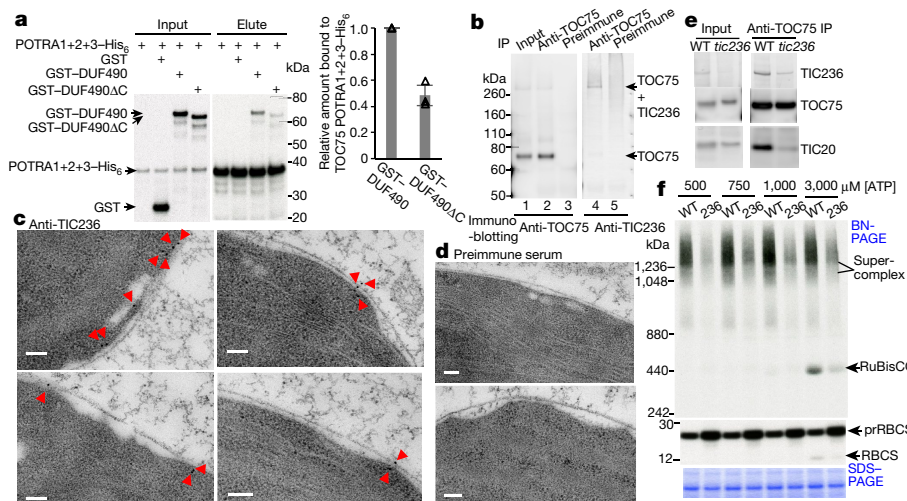


Fig. 3 | TIC236 directly binds to TOC75 and *tic236* mutant chloroplasts have reduced amounts of TOC-TIC supercomplexes. **a**, Proteins were synthesized by in vitro translation. Equal moles of GST, GST-DUF490 or GST-DUF490ΔC were incubated with a quarter amount of POTRA1-POTRA2-POTRA3-His₆ (POTRA1+2+3-His₆). Proteins pulled down by metal affinity resin were analysed by SDS-PAGE and fluorography. The amounts of GST-DUF490 and GST-DUF490ΔC pulled down were quantified. **b**, Membranes from *Arabidopsis* chloroplasts treated with 0.5 mM SMCC were solubilized by 1% LDS (input) and immunoprecipitated with anti-TOC75 or the preimmune serum, analysed by SDS-PAGE and immunoblotting, and then hybridized to anti-TOC75 or anti-TIC236 antibodies. **c**, **d**, Immunogold electron microscopy analyses of TIC236 localization. Ultra-thin sections of pea chloroplasts

were hybridized with antibodies against pea TIC236 (**c**) or the preimmune serum (**d**), and then with colloidal gold-conjugated secondary antibodies. Gold particles are indicated with red arrowheads. Scale bar, 100 nm. **e**, Total membranes of isolated *Arabidopsis* wild-type (WT) and *tic236*-mutant chloroplasts were solubilized (input) and immunoprecipitated with anti-TOC75 antibodies. **f**, ATP-depleted [³⁵S]prRBCS was incubated with ATP-depleted wild-type (WT) and *tic236* (236)-mutant chloroplasts under various concentrations of added ATP in the dark for 10 min. Chloroplasts were solubilized with 1% digitonin and analysed by BN-PAGE or SDS-PAGE. Data shown as mean \pm s.d. of three independent experiments (**a**) and representative of three (**a**, **c**, **d**, **f**) and two (**b**, **e**) independent experiments.

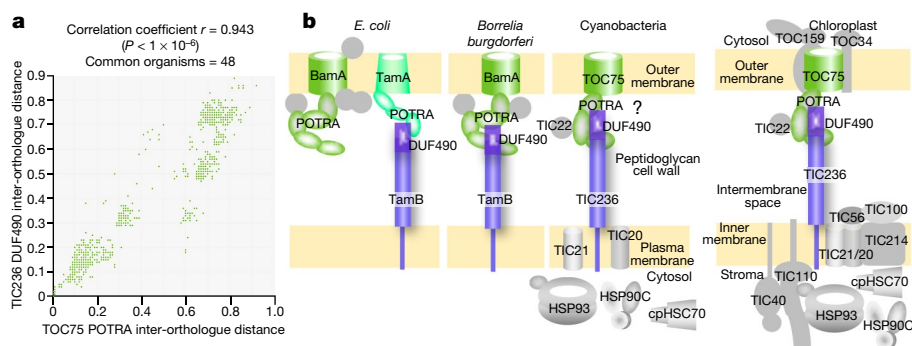


Fig. 4 | Evolutionary relationship of TIC236 and TOC75.

a, Co-evolution of DUF490 domain of TIC236 and POTRA domains of TOC75. r , Pearson correlation coefficient. One-sided t -test (see Methods for notes on P value). **b**, The TamB–TamA and TamB–BamA systems in *E. coli* and *Borrelia burgdorferi*, and the TIC236–TOC75 system in cyanobacteria and chloroplasts. Major chloroplast translocon components

wild-type chloroplasts (Fig. 3f). When the same samples were analysed by SDS–PAGE, the amount of prRBCs that was associated with the mutant chloroplasts was higher than that associated with the wild-type chloroplasts. This indicates that some prRBCs accumulated on the surface of mutant chloroplasts but could not associate with the TOC–TIC supercomplexes, and suggests that lower amounts of TOC–TIC supercomplexes were available for preprotein binding in the mutant. We next used imported [35 S]TOC34 to reflect the amounts of TOC–TIC supercomplexes, the results of which also support the observation that lower amounts of supercomplexes were present in the *tic236*-mutant chloroplasts (Extended Data Fig. 7c).

TIC236 and *E. coli* TamB⁵ have a similar domain structure and membrane topology, and both proteins directly interact with the POTRA domains of an Omp85 family member. These similarities prompted us to investigate the evolutionary relationship between TIC236 and TOC75. We first analysed the lineage distribution of TIC236. Despite the widespread presence of TamB in Gram-negative bacteria¹⁵, TamB and TIC236 orthologues are absent from fungal and animal lineages. Among eukaryotes they are only found in Rhodophyta (red algae) and Viridiplantae⁸ (green algae and land plants) and are present in all major branches of these plastid-containing organisms (Extended Data Fig. 8a, also see Methods), similar to the evolutionary path of TOC75 orthologues. We next analysed whether TIC236 and TOC75 have co-evolved through the plant lineages, which would provide further support for their functional interaction. Sequences of DUF490 domains of TIC236 orthologues and POTRA domains of TOC75 orthologues from 48 representative species (Extended Data Table 1 and Supplementary Discussion)—including cyanobacteria, red algae, green algae and land plants—were aligned and input into the MirrorTree server. A strong correlation (defined as a correlation coefficient, $r > 0.800$) was generated ($r = 0.943$, $P < 10^{-6}$) (Fig. 4a), which suggests that the DUF490 and POTRA domains co-evolved from cyanobacteria to land plants. This implies that the backbone of the chloroplast protein-import machinery evolved from a possible orthologous TIC236–TOC75 complex in cyanobacteria, and that it is related to the TamB–TamA and TamB–BamA protein secretion systems of proteobacteria (Fig. 4b). Other components of the chloroplast import machinery—such as TIC channel proteins and stromal chaperon ATPases—were recruited later and new components—such as TIC159 and TIC110—evolved, resulting in the establishment of the current chloroplast translocon.

A crystal structure of the N-terminal half of the *E. coli* TamB DUF490 domain revealed a concave, taco-shaped β -sheet with a hydrophobic interior¹⁶. Sequence analyses predict that this structure may be shared by a large portion of TamB, allowing TamB to form a chaperoning conduit¹⁶. Secondary structure prediction of TIC236 shows that it too has a high propensity for forming β -strands throughout almost

its entire length (Extended Data Fig. 2). POTRA domains of TOC75 have also been shown to have chaperone activity^{17,18}. Therefore, TIC236 together with the POTRA domains of TOC75 may provide a continuous chaperoned passage for preprotein translocation across the intermembrane space (see Supplementary Discussion). Inward transport may be ensured by other mechanisms, such as increased transit-peptide binding affinity¹⁹ or pulling by stromal ATPase motors^{20,21}.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0713-y>.

Received: 26 March 2018; Accepted: 18 October 2018;

Published online 21 November 2018.

- Shi, L. X. & Theg, S. M. The chloroplast protein import system: from algae to trees. *Biochim. Biophys. Acta* **1833**, 314–331 (2013).
- Paila, Y. D., Richardson, L. G. L. & Schnell, D. J. New insights into the mechanism of chloroplast protein import and its integration with protein quality control, organelle biogenesis and development. *J. Mol. Biol.* **427**, 1038–1060 (2015).
- Geissler, A. et al. The mitochondrial presequence translocase: an essential role of Tim50 in directing preproteins to the import channel. *Cell* **111**, 507–518 (2002).
- Yamamoto, H. et al. Tim50 is a subunit of the TIM23 complex that links protein translocation across the outer and inner mitochondrial membranes. *Cell* **111**, 519–528 (2002).
- Selkrig, J. et al. Discovery of an archetypal protein transport system in bacterial outer membranes. *Nat. Struct. Mol. Biol.* **19**, 506–510 (2012).
- Chu, C. C. & Li, H.-m. Protein import into isolated pea root leucoplasts. *Front. Plant Sci.* **6**, 690 (2015).
- Tzafirir, I. et al. Identification of genes required for embryo development in *Arabidopsis*. *Plant Physiol.* **135**, 1206–1220 (2004).
- Matsushima, R. et al. Amyloplast-localized SUBSTANDARD STARCH GRAIN4 protein influences the size of starch grains in rice endosperm. *Plant Physiol.* **164**, 623–636 (2014).
- Chen, Y. L., Chen, L. J. & Li, H.-m. Polypeptide transport-associated domains of the Toc75 channel protein are located in the intermembrane space of chloroplasts. *Plant Physiol.* **172**, 235–243 (2016).
- Heinz, E. & Lithgow, T. A comprehensive analysis of the Omp85/TpsB protein superfamily structural diversity, taxonomic occurrence, and evolution. *Front. Microbiol.* **5**, 370 (2014).
- Iqbal, H., Kennedy, M. R., Lybecker, M. & Akins, D. R. The TamB ortholog of *Borrelia burgdorferi* interacts with the β -barrel assembly machine (BAM) complex protein BamA. *Mol. Microbiol.* **102**, 757–774 (2016).
- Chen, L. J. & Li, H.-m. Stable megadalton TOC–TIC supercomplexes as major mediators of protein import into chloroplasts. *Plant J.* **92**, 178–188 (2017).
- Tu, S. L. et al. Import pathways of chloroplast interior proteins and the outer-membrane protein OEP14 converge at Toc75. *Plant Cell* **16**, 2078–2088 (2004).
- Li, H.-m. & Chen, L.-J. A novel chloroplastic outer membrane-targeting signal that functions at both termini of passenger polypeptides. *J. Biol. Chem.* **272**, 10968–10974 (1997).
- Heinz, E., Selkrig, J., Belousoff, M. J. & Lithgow, T. Evolution of the Translocation and Assembly Module (TAM). *Genome Biol. Evol.* **7**, 1628–1643 (2015).
- Josts, I. et al. The structure of a conserved domain of TamB reveals a hydrophobic β taco fold. *Structure* **25**, 1898–1906.e5 (2017).

17. Paila, Y. D. et al. Multi-functional roles for the polypeptide transport associated domains of Toc75 in chloroplast protein import. *eLife* **5**, e12631 (2016).
18. O'Neil, P. K. et al. The POTRA domains of Toc75 exhibit chaperone-like function to facilitate import into chloroplasts. *Proc. Natl Acad. Sci. USA* **114**, E4868–E4876 (2017).
19. Komiya, T. et al. Interaction of mitochondrial targeting signals with acidic receptor domains along the protein import pathway: evidence for the 'acid chain' hypothesis. *EMBO J.* **17**, 3886–3898 (1998).
20. Liu, L., McNeillage, R. T., Shi, L. X. & Theg, S. M. ATP requirement for chloroplast protein import is set by the K_m for ATP hydrolysis of stromal Hsp70 in *Physcomitrella patens*. *Plant Cell* **26**, 1246–1255 (2014).
21. Huang, P. K., Chan, P. T., Su, P. H., Chen, L. J. & Li, H.-m. Chloroplast Hsp93 directly binds to transit peptides at an early stage of the preprotein import process. *Plant Physiol.* **170**, 857–866 (2016).

Acknowledgements We thank Y.-S. Teng for technical assistance at the initial stage of this project, M. Akita for the *Cyanidioschyzon merolae* 10D genomic DNA, the proteomics and imaging cores of the Institute of Molecular Biology and the proteomics core of the Institute of Biological Chemistry of Academia Sinica for technical assistance, ABRC for the *tic236-1* and *tic236-2* mutant seeds and the plant genome project of RIKEN Genomic Sciences Center for *tic236-3* mutant seeds. This work was supported by the Ministry of Science and Technology (MOST 107-2321-B-001-001) and Academia Sinica of Taiwan (to H.-m.L.).

Reviewer information Nature thanks N. Pfanner, D. Schnell and S. Theg for their contribution to the peer review of this work.

Author contributions Y.-L.C., L.-J.C. and C.-C.C. performed crosslinking and immunoprecipitations; Y.-L.C. performed subplastid fractionation, electron microscopy and in vitro pull down; L.-J.C. and C.-C.C. performed alkaline extractions; L.-J.C. performed imports, mutant isolation and BN-PAGE; L.-J.C. and H.-m.L. performed sucrose density fractionations; P.-K.H. performed evolutionary analyses; J.-R.W. determined the length of TIC236 transit peptide; and H.-m.L. wrote the paper.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0713-y>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0713-y>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to H.-m.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Chloroplast isolation, fractionation, protein import, and protease and alkaline extraction treatments. Pea (*Pisum sativum* 'Green Arrow', 7–11-day-old) chloroplast isolation²², growth of *Arabidopsis* (14 or 21 days on MS agar medium with 2% sucrose), *Arabidopsis* chloroplast isolation, chloroplast fractionation after hypertonic lysis and alkaline extraction (0.1 M Na₂CO₃, pH 11.5), import of [³⁵S] Met-labelled preproteins into isolated chloroplasts²³, thermolysin and trypsin treatments of isolated chloroplasts⁹ and leucoplast isolation from roots of 4-day-old pea seedlings grown in the dark⁶ were performed as previously described. In Fig. 1b, each lane was loaded with 2.0 µg proteins for the gel analysing TIC236, and 0.5 µg for gels analysing all other proteins. In Fig. 1c, proteins derived from the same number of chloroplasts were loaded in the two lanes and CAB and RBCS were used as controls for integral-membrane and soluble proteins, respectively. In Fig. 1d, because thermolysin cannot penetrate the outer membrane and trypsin can penetrate the outer but not inner membrane²⁴, treatments with these two proteases were used to distinguish on which side of the inner membrane the protein resides. The amounts of TIC110, TOC75 and TOC34 were analysed to serve as controls for effectiveness and proper quenching of proteases. Isolated pea chloroplasts were treated with 200 µg/ml thermolysin or 200 µg/ml trypsin. The same amount of proteins was loaded in each lane. For all import experiments, intact chloroplasts were re-isolated after incubating with [³⁵S]-labelled preproteins under indicated conditions and analysed by SDS–PAGE. The gels were stained with Coomassie blue and dried for fluorography.

[³⁵S]Met-labelled preproteins were in vitro-transcribed and translated using the TNT coupled wheat germ or reticulocyte lysate system, and SP6 or T7 RNA polymerase (Promega). For the construct that expressed EMB2410(1–227), the coding region for residues 1 to 405 of *Arabidopsis* TIC236 was amplified by PCR and cloned into pSP72 under the direction of the SP6 promoter. Residue 228 was then mutated from Leu to a stop codon. For estimation of the length of the transit peptide, the initiation Met was first mutated to Ile and residues 30, 35 38 and 46 were further mutated individually to Met. Site-directed mutagenesis was performed using the QuikChange Lightning Site-Directed Mutagenesis kit (Agilent). Accession numbers or references for other preproteins used are: prRBCS²⁵, prOE23 (At1g06680), prHSP93 (L09547), prTIC40 (AY157668), TOC34 (At5g05000¹⁴), and OEP14²⁶. For import time-course and regular import experiments, such as those shown in Fig. 2d–f and Extended Data Fig. 6b, preproteins were incubated with chloroplasts under 1 mM ATP in the light at room temperature for 10 min, or for the amount of time indicated. For quantification, the amount of imported mature proteins (mature plus intermediate proteins for TIC40) were quantified by phosphorimager, corrected for loading by quantifying the amount of RBCL (for prRBCS and prHSP93 import) or CAB (for prTIC40 import), and then normalized to the amount imported in the wild-type chloroplasts at 20 min. For TOC34 import shown in Extended Data Fig. 7c, chloroplasts were incubated with [³⁵S]TOC34 under 3 mM ATP at room temperature for 25 min. For binding and import experiments shown in Fig. 3f, chloroplasts were depleted of internal ATP by keeping them at 4°C in the dark for 2 h. [³⁵S]prRBCS was filtrated to remove nucleotides and then incubated with the energy-depleted chloroplasts under green safe light with increasing concentrations of exogenous ATP at room temperature for 10 min. Re-isolated intact chloroplasts were solubilized by 1% digitonin¹².

Crosslinking, immunoprecipitation and antibody preparations. Immunoprecipitation of solubilized chloroplasts after DSP or SMCC crosslinking was performed as described²¹. In brief, for experiments shown in Figs. 1e, 3e, chloroplasts were treated with 1 mM DSP for 15 min in the dark at 4°C, quenched with 100 mM glycine, and lysed by freeze–thaw cycles. The lysate was centrifuged at 3,000g for 10 min at 4°C to collect the thylakoid membranes. Then, the supernatant was centrifuged at 100,000g for 45 min at 4°C to collect the envelope-membrane fraction. The two membrane fractions were solubilized separately with 1% decylmaltoside. The two fractions were either combined, or only the envelope fraction was used. The solubilized membranes were clarified by centrifuging at 100,000g for 15 to 30 min. The supernatant was designated as total 'input' for immunoprecipitation. Antibodies against the POTRA domains of *Arabidopsis* TOC75⁹, the soluble domain of *Arabidopsis* TIC110¹² or the DUF490 of pea TIC236, and then protein A–agarose beads (Pierce) were added. The beads were washed in co-immunoprecipitation buffer (50 mM HEPES–NaOH, pH 7.5, 150 mM NaCl, 4 mM MgCl₂, 10% glycerol and 1× EDTA-free protease inhibitor cocktail (Roche)) containing 1% decylmaltoside and eluted using 2× SDS–PAGE sample buffer. For the anti-TOC75 immunoprecipitation shown in Fig. 1e, the two lanes were analysed on the same blotting membrane and developed for the same amount of time (some intervening lanes have been removed). For the anti-TIC236 immunoprecipitation shown in Figs. 1e, 2.5-fold immunoprecipitates were used for the analyses of TIC20 and IEP37, compared to analyses of other proteins. SMCC is a heterobifunctional

(cysteine to primary amine) crosslinker. For SMCC crosslinking, chloroplasts were treated with 0.5 mM SMCC for 30 min in the dark at 4°C and quenched with 50 mM DTT. Membranes from lysed chloroplasts were solubilized with 1% LDS. The solubilized membranes were diluted with nine volumes of buffer without LDS before being used for immunoprecipitation. Washings were performed in co-immunoprecipitation buffer containing 0.1% LDS. Other procedures are the same as described for DSP crosslinking.

Rabbit antibodies were generated against the DUF490 domain of pea (residues 1794 to 2211) and *Arabidopsis* (residues 1751 to 2166) TIC236. The corresponding cDNA was amplified by PCR using pea and *Arabidopsis* first-strand cDNA as templates, and cloned into the XhoI/PstI sites of pRSET-B with an N-terminal His₆ tag. Recombinant His₆–DUF490 proteins were purified from the pellet fraction of *E. coli* lysate after solubilization in 8 M urea and purified by TALON metal-affinity resins. Mouse antibodies against TIC236 were generated against a peptide corresponding to *Arabidopsis* TIC236 residues 1957 to 1988 (sequence VNLVATQVRLKREHLNVAKFEPHGLDPLDL, Extended Data Fig. 2). The specificity of these three antibodies is shown in Extended Data Fig. 8b. The mouse anti-DUF490-peptide antibodies were used in Figs. 1b–e, 3e (TOC75 immunoprecipitation) and Extended Data Fig. 5; the rabbit anti-*Arabidopsis*-DUF490 antibodies were used in Fig. 3b; and the rabbit anti-pea-DUF490 antibodies were used in Figs. 1e–g, 3c, e (TIC110 and TIC236 immunoprecipitation) and Extended Data Fig. 7b. The rabbit anti-IEP37 antibody was generated against residues 81–281 of *Arabidopsis* IEP37 (At3g63410). A cDNA fragment encoding the corresponding region was amplified by PCR and subcloned into the BamHI/EcoRI site of pRSET-A with an N-terminal His₆ tag. Antibodies against TIC110, TOC34 and pea TOC75¹³, *Arabidopsis* TOC75⁹, TIC40²⁷ and *Arabidopsis* TIC201 (At1g04940)²⁸ have previously been described. The following antibodies were purchased from Agrisera: cytosolic sucrose phosphate synthase (SPS) (AS03 035A); cpHSC70 (AS08 348); GS2 (AS08 296); and CAB (AS01 004). For detection of proteins in immunoprecipitates on immunoblots, EasyBlot secondary antibodies (anti-native rabbit IgG, GeneTex) were used to decrease interference from denatured IgG heavy and light chains.

Sucrose density gradient centrifugation, immunogold electron microscopy and BN-PAGE. For sucrose density gradient centrifugation, [³⁵S]prRBCS was imported into pea chloroplasts under 3 mM ATP for 2 min at room temperature. Re-isolated intact chloroplasts were solubilized by 1% decylmaltoside in BN buffer (50 mM Bis–Tris, pH 7.2, 50 mM NaCl, and 1× EDTA-free protease inhibitor cocktail (Roche)) and 10% glycerol at 4°C for 30 min at a chlorophyll concentration of 2 mg/ml. Insoluble materials were removed by centrifugation at 100,000g at 4°C for 15 min. The supernatant was layered onto a 15 to 45% linear sucrose gradient prepared in BN buffer and 0.3% decylmaltoside and the gradient was centrifuged in a SW41 rotor at 30,000 rpm for 18 h at 4°C. Fractions were retrieved from the top of the gradient and aliquots were either directly analysed or concentrated by chloroform–methanol precipitation, and analysed by SDS–PAGE and immunoblotting or fluorography.

For immunogold electron microscopy, isolated pea chloroplasts were incubated in import buffer (330 mM sorbitol and 50 mM HEPES–KOH, pH 8.0) containing 0.6 M sucrose for 10 min on ice to increase the distance between the two envelope membranes²⁹. The chloroplasts were pelleted down and processed for high-pressure freezing, freeze substitution, embedding, ultra-thin sectioning, immunogold labelling and image acquisition by electron microscopy as previously described⁹. Affinity-purified antibodies against the DUF490 domain of pea TIC236 and the preimmune serum IgG of the same rabbit and of the same IgG concentration were used for labelling. Ribosomes in chloroplasts have a similar size to the gold particles but are not as dark or as rounded as the gold particles. For Fig. 3c, d, antiserum dilution was 1:125 for the top row and 1:200 for the bottom row.

BN-PAGE analyses of chloroplasts after import reactions were performed as described¹². Chloroplasts used for BN-PAGE analyses were isolated in the presence of 1× EDTA-free protease inhibitor cocktail (Roche).

Expression of *Arabidopsis* TIC236 and identification of *tic236* mutants. Overviews of TIC236 (At2g25660) RNA expression in various tissues were obtained from the *Arabidopsis* eFP browser (<http://bbc.botany.utoronto.ca/efp/cgi-bin/efpWeb.cgi>). The *tic236-1* (SALK_048770³⁰, T-DNA insertion at base pair 4966–4967 of the tenth exon; the translation initiation site is designated as +1 and the 5' UTR is represented by negative numbers) and *tic236-2* (SAIL104-F07, T-DNA insertion at –306 and –307) mutants, both in the *Col* ecotype, were obtained from the *Arabidopsis* Biological Resource Center (ABRC). The *tic236-3* mutant (RIKEN PST00216^{31,32}) in the *No-0* ecotype was obtained from RIKEN BRC Experimental Plant Division. Two lines (PST00216 and PST01662) were obtained of this latter mutant. They have a Ds insertion in the same position in the 5' UTR at –276 and –277. If not specified otherwise, for all experiments the *tic236-2* allele was used and *Col* represents the wild-type control.

Full-length TIC236 cDNA could be amplified from *Arabidopsis* leaf first-strand cDNA by PCR but could not be cloned into an *E. coli* vector, similar to the

difficulties encountered with rice *SSG4* cDNA⁸. Clones consistently had deletions in a region of the N-terminal half. Therefore, we designed primers to mutate the DNA sequence in that region but preserve the amino acid sequence. Finally, a clone that encoded a polypeptide with the same amino acid sequence as that published in the *Arabidopsis* Information Resource (TAIR) was obtained, but this cDNA could still only be maintained in a low-copy-number plasmid. A construct of the CaMV 35S promoter driving expression of this cDNA fused to a YC tag (the C-terminal one-third of YFP) could not complement the *tic236-2* mutant, but no transgenic protein expression could be detected either. In light of the fact that the promoter region of rice *SSG4* is also toxic to *E. coli*⁸ and the fact that we have two alleles in different ecotypes showing the same phenotypes, no further attempts at complementation were made.

In vitro pull-down assays. To express POTRA1–POTRA2–POTRA3–His₆, the coding region for *Arabidopsis* TOC75 residues 141 to 468 was amplified by PCR and cloned into pET-22b, resulting in a C-terminal His₆ tag⁹. For POTRA1–POTRA2–His₆ and POTRA1–His₆, primers listed in the Extended Data Table 2 and the QuikChange Lightning Site-Directed Mutagenesis kit (Agilent) were used to delete POTRA3 and POTRA2–POTRA3 from the POTRA1–POTRA2–POTRA3–His₆ construct. To express GST–DUF490 and GST–DUF490ΔC, the coding region for GST was excised from pGEX-5X-1 by HincII and cloned into the EcoRV site of pBluscriptSK. The resulting plasmid was cut with XhoI and XbaI, and the insert was cloned into the XhoI/XbaI site of pSP72, creating the pSP72–GST plasmid. The coding regions for *Arabidopsis* TIC236 residues 1752–2166 and 1752–2150 were amplified by PCR and cloned into the EcoRI/SpeI site of pSP72–GST for the production of GST–DUF490 and GST–DUF490ΔC, respectively. [³⁵S]Met-labelled proteins were synthesized by *in vitro* transcription and translation and first analysed by SDS–PAGE and phosphorimager to measure the relative amount of protein synthesized. After normalization to the number of methionine residues in each protein (9, 20, 20, 9, 9 and 4 for GST, GST–DUF490, GST–DUF490ΔC, POTRA1–POTRA2–POTRA3–His₆, POTRA1–POTRA2–His₆ and POTRA1–His₆, respectively), equal moles of GST, GST–DUF490 or GST–DUF490ΔC were incubated with a quarter that amount of moles of POTRA1–POTRA2–POTRA3–His₆, POTRA1–POTRA2–His₆ or POTRA1–His₆ for 30 min at room temperature. TALON metal-affinity resin (Clontech) was then added to pull down the POTRA proteins and associated proteins. Resin was washed 3 times with washing buffer (50 mM HEPES–NaOH, pH 7.5, 200 mM NaCl) and eluted with 2× SDS–PAGE sample buffer, analysed by SDS–PAGE and quantified by phosphorimager. For Fig. 3a and Extended Data Fig. 7a, equal moles of POTRA proteins were loaded among the lanes of each gel.

Evolutionary analyses. Lineage distribution of TamB and TIC236 was first re-analysed using the sequence of the DUF490 of *Arabidopsis* TIC236 to search the InterPro 57.0 database released in April 2016³³. A previous study⁸ that used the DUF490 of rice *SSG4* as a query reported that proteins with a DUF490 domain were found in bacteria and the Viridiplantae (green algae and land plants) but not in animals⁸. We reached a similar conclusion. In the InterPro 57.0 database, there are 6,549 proteins annotated as having a DUF490 domain: 245 in cyanobacteria, 6,188 in other bacteria, 113 in Viridiplantae and one in Rhodophyta (red algae). Although two metazoan sequences are annotated as having a DUF490 domain—A0A077ZFV6 in *Trichuris trichiura* (whipworm) and E9HXV0 in *Daphnia pulex* (water flea)—these annotations most probably resulted from sample contamination. A0A077ZFV6 is a fusion protein containing an almost complete TamA in the N terminus and an almost complete TamB in the C terminus, and the sequences are 99% identical to *E. coli* TamA and TamB, respectively. Similarly, E9HXV0 also exhibits 72% identity with the C terminus of TamB of *Pseudomonas taeanensis*. We did not find any other DUF490-containing proteins in any other Arthropoda. Therefore, it is most likely that both proteins are the result of sample contamination—especially because the whipworm infects large intestines and the water flea feeds on bacteria. Furthermore, the status of both A0A077ZFV6 and E9HXV0 is listed as ‘unreviewed’ in the TrEMBL section of UniProtKB.

To cover all major branches of the Viridiplantae for phylogenetic relationship analyses, sequences containing DUF490 were searched with BLASTP in the GenBank, Phytozome and the Spruce genome project (www.congenie.org) using *Arabidopsis* TIC236 as the query sequence. The putative orthologues of TIC236 were predicted through reciprocal best hits with BLASTP. We further assembled the TIC236 homologue in another red alga, *C. merolae* strain 10D, by re-sequencing the corresponding region. The coding region for a possible TIC236 homologue in *C. merolae* strain 10D is currently annotated as two separate records in its genome project: CMO228C in ‘*C. merolae* annotated CDS’ and ER037014 in ‘*C. merolae* full-length cDNA reverse’ (EST database). We re-sequenced the region and found that an erroneous extra G residue has been inserted at the C terminus of CMO228C. After removing that extra base, the two separate genes become one open reading frame (ORF). We have submitted the ORF under accession GenBank: MG799551. The bacterial proteins containing the DUF490 domain were acquired

using InterPro. Multiple sequence alignments were performed using MAFFT (version 7, web server; choosing L-INS-i and all other settings as default) and the domain boundaries were determined. The sequence alignments corresponding to residues 1755–2166 (the DUF490 domain) of *Arabidopsis* TIC236 were extracted for further domain alignment using MAFFT with G-INS-i and all other settings as default. Neighbour-joining trees were constructed in MEGA 7.0.26³⁴ using the Jones–Thompson–Taylor model and pairwise deletion for gap filling. To test inferred phylogenies, we used bootstraps with 1,000 bootstrap replicates.

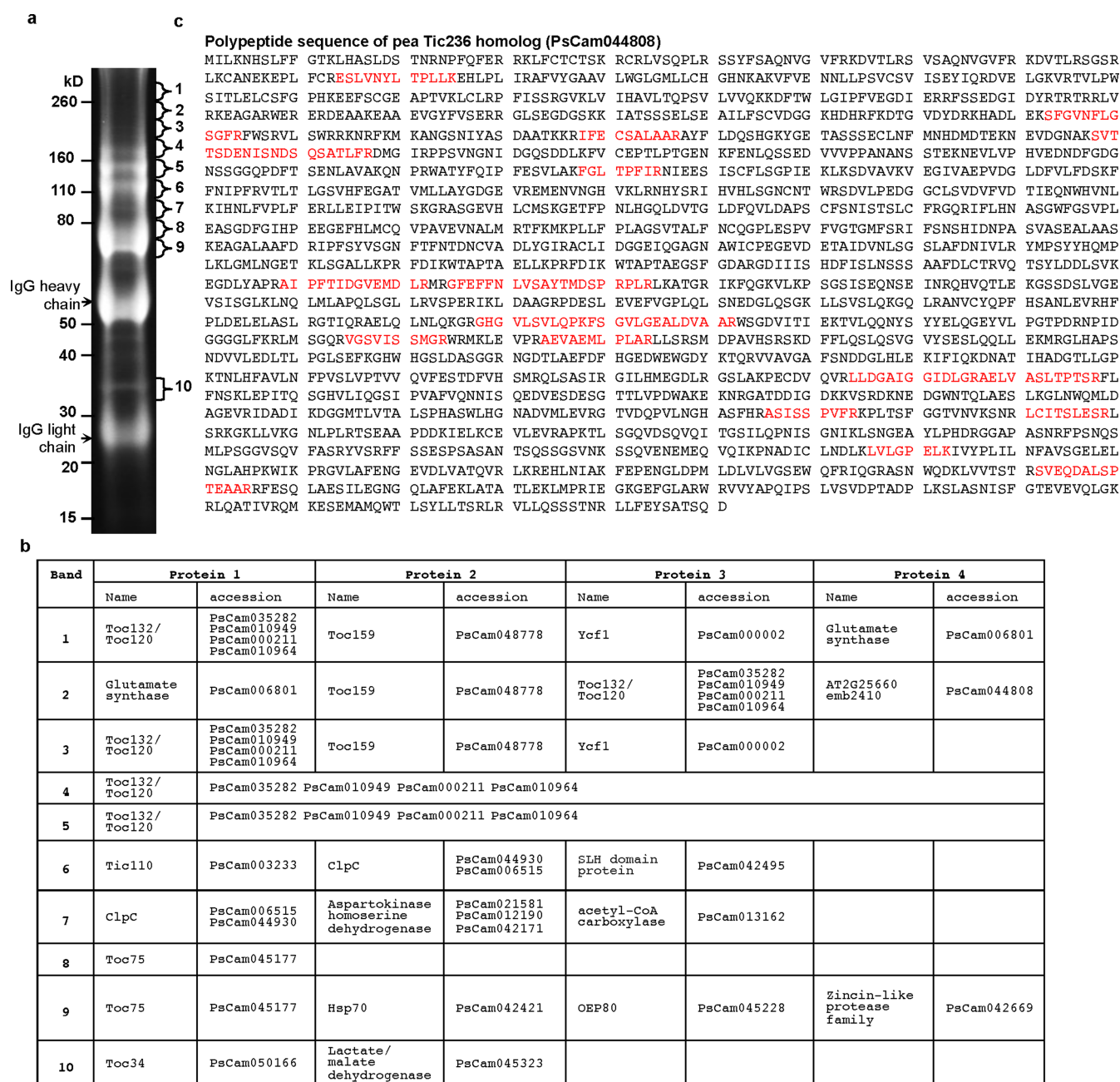
Co-evolution is believed to maintain the specific interactions of interacting proteins, and can be reflected in the similarity of the corresponding phylogenetic trees of the proteins³⁵. We used the MirrorTree method³⁶, which quantifies tree similarity by calculating the Pearson correlation coefficient between the distance matrices underlying the trees. Putative orthologues of TOC75 and TIC236 (except those from *Nitella mirabilis*) were identified in GenBank and Phytozome by reciprocal BLASTP using amino acid sequences of *Arabidopsis thaliana* TOC75-III (At3g46740) and TIC236 (At2g25660) as queries. For *N. mirabilis*, TOC75 and TIC236 sequences were obtained from the Transcriptome Shotgun Assembly Sequence Database using TBLASTN. Multiple sequence alignments were performed using MAFFT as described above. Residues corresponding to 141–468 of TOC75-III and 1755–2166 of TIC236 were defined as the POTRA and DUF490 domains, respectively. Multiple sequence alignment results were used as input for the MirrorTree web server³⁷ to construct phylogenetic trees, define evolutionary distances and calculate the correlation coefficient between the evolutionary distances of each tree. The plot in Fig. 4a is a simplified representation of the correlation between the inter-protein distances calculated for two phylogenetic trees: one tree for the DUF490 domains of TIC236, and the other for the POTRA domains of TOC75. The *P* value in this case is so small that the MirrorTree Server can only return *P* < 10^{−6}. We thus report the *t* value and sample size (*n*) instead: the *t* value is 95.08392, and *n* is 1,128. We also used the TDist function in R program for calculating the precise value, but it returned *P* = 0 because the *P* is too small.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All data generated or analysed during this study are included in the article and its Supplementary Information. Full gel blots can be found in Supplementary Fig. 1. Any other data are available from the corresponding author upon reasonable request.

- Perry, S. E., Li, H.-m. & Keegstra, K. In vitro reconstitution of protein transport into chloroplasts. *Methods Cell Biol.* **34**, 327–344 (1991).
- Chu, C. C. & Li, H.-m. Determining the location of an *Arabidopsis* chloroplast protein using *in vitro* import followed by fractionation and alkaline extraction. *Methods Mol. Biol.* **774**, 339–350 (2011).
- Cline, K., Werner-Washburne, M., Andrews, J. & Keegstra, K. Thermolysin is a suitable protease for probing the surface of intact pea chloroplasts. *Plant Physiol.* **75**, 675–678 (1984).
- Lubben, T. H. & Keegstra, K. Efficient *in vitro* import of a cytosolic heat shock protein into pea chloroplasts. *Proc. Natl Acad. Sci. USA* **83**, 5502–5506 (1986).
- Tu, S.-L. & Li, H.-m. Insertion of OEP14 into the outer envelope membrane is mediated by proteinaceous components of chloroplasts. *Plant Cell* **12**, 1951–1960 (2000).
- Chou, M. L., Chu, C. C., Chen, L. J., Akita, M. & Li, H.-m. Stimulation of transit-peptide release and ATP hydrolysis by a cochaperone during protein import into chloroplasts. *J. Cell Biol.* **175**, 893–900 (2006).
- Teng, Y. S. et al. Tic21 is an essential translocon component for protein translocation across the chloroplast inner envelope membrane. *Plant Cell* **18**, 2247–2257 (2006).
- Keegstra, K. & Yousif, A. E. Isolation and characterization of chloroplast envelope membranes. *Methods Enzymol.* **118**, 316–325 (1986).
- Alonso, J. M. et al. Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* **301**, 653–657 (2003).
- Ito, T. et al. A new resource of locally transposed *Dissociation* elements for screening gene-knockout lines in silico on the *Arabidopsis* genome. *Plant Physiol.* **129**, 1695–1699 (2002).
- Kuromori, T. et al. A collection of 11 800 single-copy *Ds* transposon insertion lines in *Arabidopsis*. *Plant J.* **37**, 897–905 (2004).
- Finn, R. D. et al. InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Res.* **45**, D190–D199 (2017).
- Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).
- Ochoa, D. & Pazos, F. Practical aspects of protein co-evolution. *Front. Cell Dev. Biol.* **2**, 14 (2014).
- Pazos, F. & Valencia, A. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng.* **14**, 609–614 (2001).
- Ochoa, D. & Pazos, F. Studying the co-evolution of protein families with the MirrorTree web server. *Bioinformatics* **26**, 1370–1371 (2010).
- Drozdetkiy, A., Cole, C., Procter, J. & Barton, G. J. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* **43**, W389–W394 (2015).



Extended Data Fig. 1 | Identification of TOC75-interacting proteins.

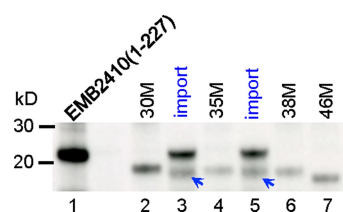
Leucoplasts were isolated from four-day-old pea roots as described⁶ and treated with 1 mM DSP. Membrane fractions were solubilized with 1% decylmaltoside and immunoprecipitated with anti-TOC75 antibody. **a**, Immunoprecipitates were analysed by SDS-PAGE and stained with SYPRO Ruby. Gel shown is a representative of two technical repeats. **b**, Gel slices as numbered in **a** were excised and processed for in-gel trypsin digestion and analysed by an LTQ-Orbitrap XL spectrometer.

The tandem mass spectrometry (MS/MS) raw data were processed by the extractmsn.exe program (Thermo) and searched against the 'Pea RNA-Seq gene atlas' (<http://bios.dijon.inra.fr/FATAL/cgi/pscam.cgi>) database using the Mascot Daemon 2.4.1 server. The top four protein hits in each band with the correct molecular mass range and more than 15 peptide matches are listed. **c**, Polypeptide sequence of pea TIC236 preprotein (PsCam044808). Peptides identified in liquid chromatography with MS/MS are marked in red.

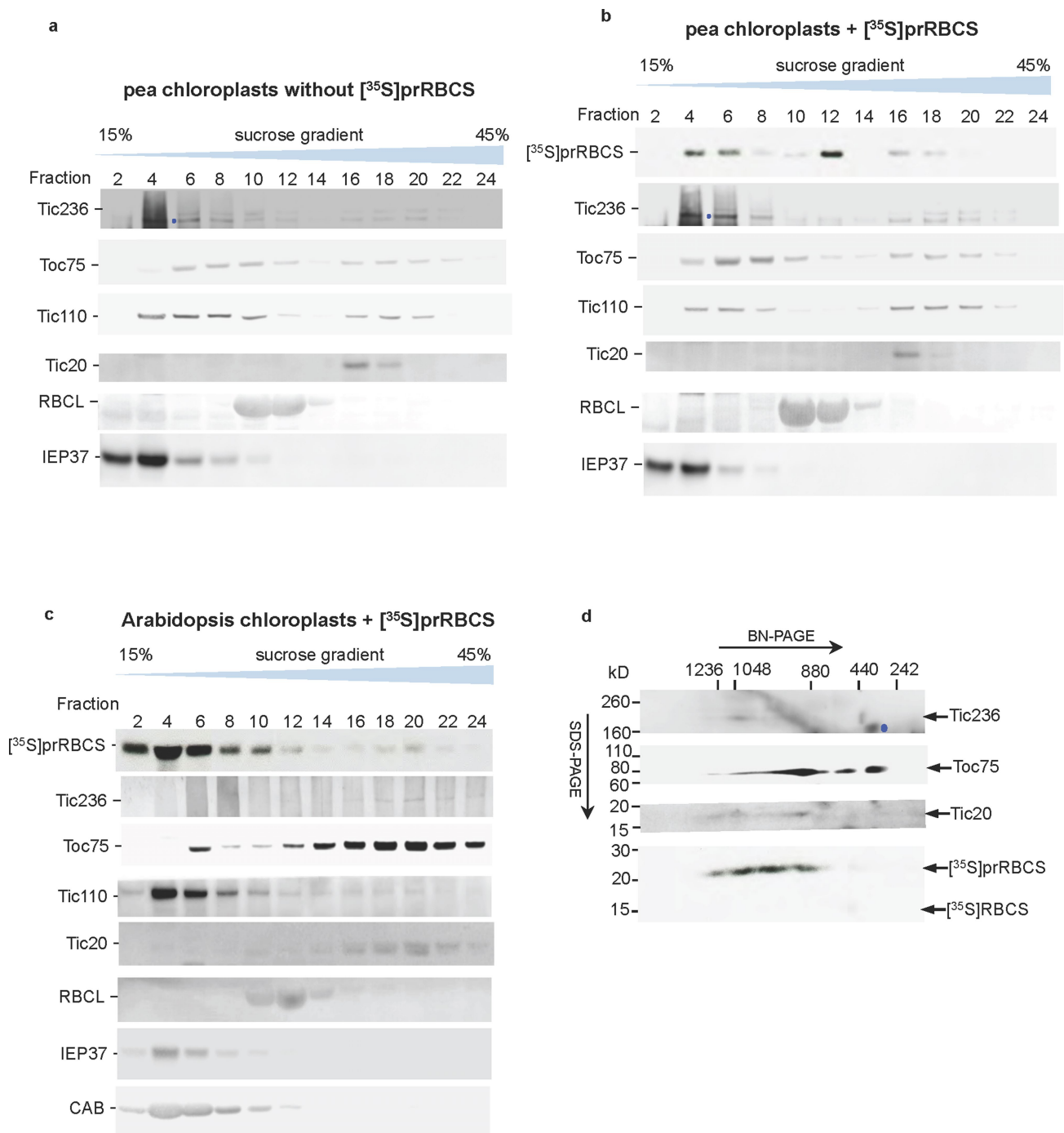
[illegible]

Extended Data Fig. 2 | Deduced polypeptide sequence of *Arabidopsis* TIC236 (At2g25660) and prediction of its secondary structure. The transit peptide is shown in green, the transmembrane domain in purple and the DUF490 domain in orange. The peptide used to generate the

mouse antibodies is underlined. Secondary-structure prediction is shown underneath the amino acid sequence and was generated using the JPred4 web server³⁸. H, α -helix; E (blue), β -strand.

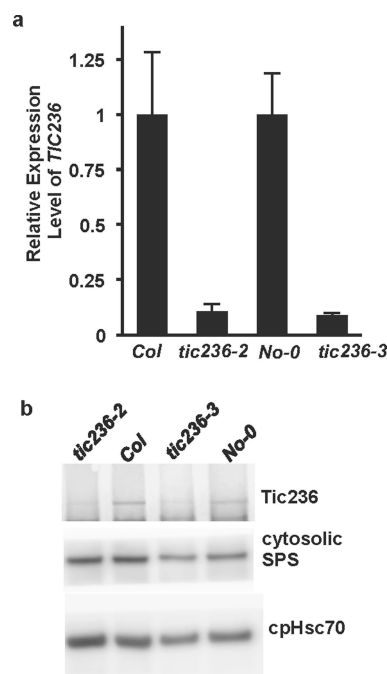


Extended Data Fig. 3 | TIC236 has a cleavable transit peptide that is approximately 37 amino acids in length. A C-terminally truncated clone that encodes residues 1 to 227 of EMB2410 was constructed. [^{35}S] EMB2410(1–227) was synthesized as a protein about 23 kDa in size (lane 1) and, when imported into chloroplasts, it was processed into a mature protein slightly smaller than 20 kDa (lanes 3 and 5, blue arrows). To estimate the processing site, a series of N-terminally truncated clones were generated by mutating the initiation methionine to isoleucine, and then individually mutating residues 30, 35, 38 or 46 to methionine (labelled 30M, 35M, 38M and 46M, respectively). In vitro translation initiated from methionine at residue 38 produced a protein of approximately the same size as the mature protein (lane 6). Thus, the transit peptide of EMB2410 is estimated to be 37 residues in length and the calculated molecular mass of EMB2410 after transit peptide removal is 236 kDa (residues 38 to 2166). Data shown are representative of two independent experiments.

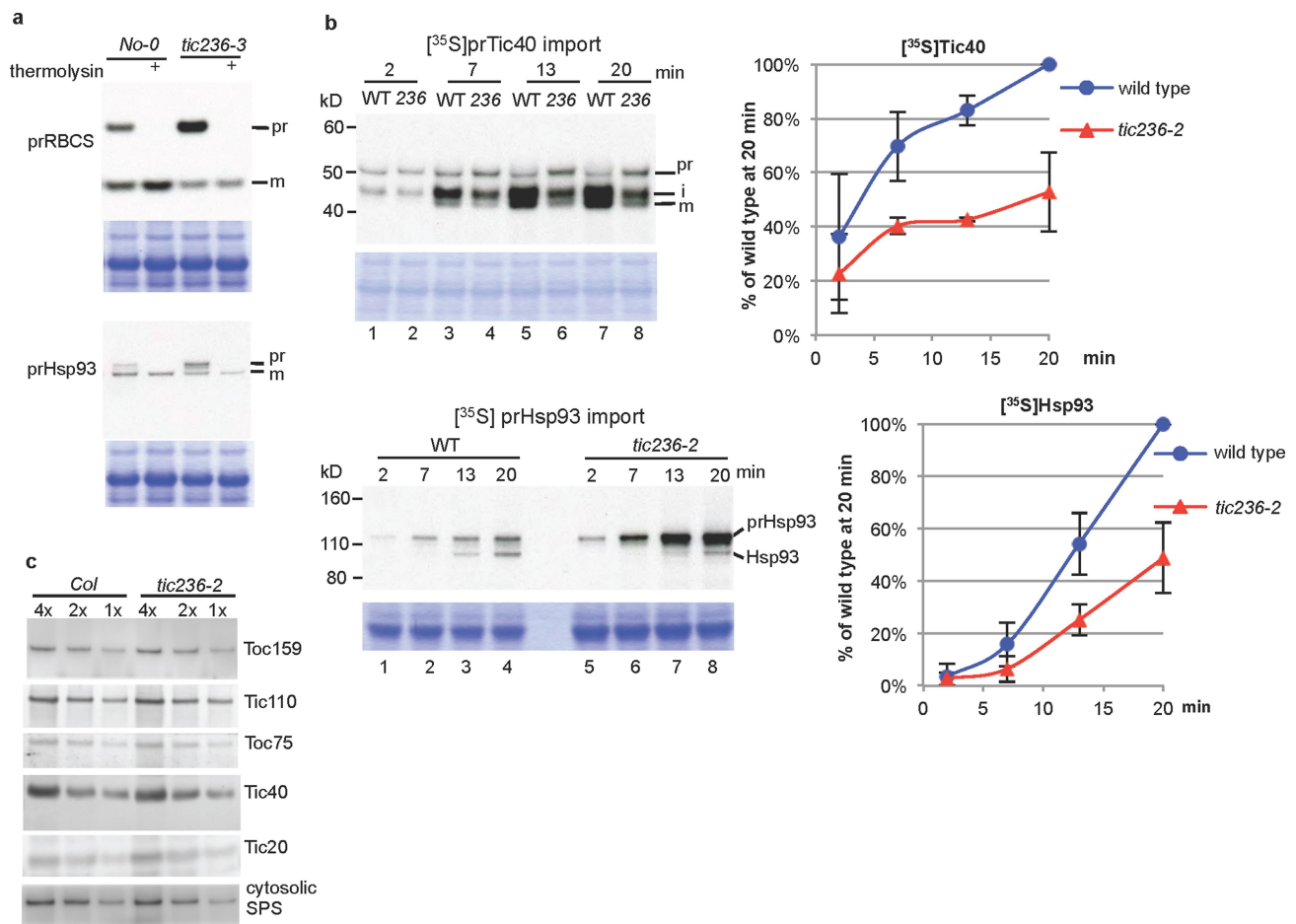


Extended Data Fig. 4 | TIC236 is a member of the TOC-TIC supercomplexes. a, b, Pea chloroplasts without (a) or with (b) (imported into chloroplasts under the conditions of 3 mM ATP at room temperature for 2 min) translocating [³⁵S]prRBCS were solubilized by 1% decylmaltoside and analysed side-by-side in 15 to 45% linear sucrose density gradients. Fractions were analysed by SDS-PAGE followed by fluorography for [³⁵S]prRBCS and immunoblotting for major translocon and control proteins. [³⁵S]prRBCS sedimented in three peaks, two of which were near the top or middle of the gradient and most probably correspond to free [³⁵S]prRBCS or [³⁵S]prRBCS that is non-specifically bound to the RuBisCO complex during solubilization. The third peak was located further into the gradient, around fractions 16 to 20. Some TOC75, TOC159 and TIC110 also sedimented in these higher-density fractions, and TIC20 was detected only in fractions 16 to 18. These results suggest that fractions 16 to 18 hosted the TOC-TIC supercomplexes.

Three control proteins (RBCL, IEP37 and CAB) were detected only at the top or centre of the gradient (see Fig. 1f and Extended Data Fig. 4c for the sedimentation pattern of CAB). The same results were observed when chloroplasts without translocating [³⁵S]prRBCS were fractionated (a). c, The same experiment as shown in b but using *Arabidopsis* chloroplasts. d, Pea chloroplasts that contain [³⁵S]prRBCS imported under 3 mM ATP at room temperature for 2 min were solubilized with 1% digitonin and analysed by 2D BN-PAGE followed by immunoblotting for TIC236, TOC75 and TIC20, or fluorography for [³⁵S]prRBCS. The blue dot indicates degraded TIC236. The result from a similar experiment is presented in Fig. 1g. Here we show that the same results were obtained either with or without translocating [³⁵S]prRBCS, and that the positions of the supercomplexes were further supported by the presence of translocating [³⁵S]prRBCS. All data shown are representative of at least two independent experiments.

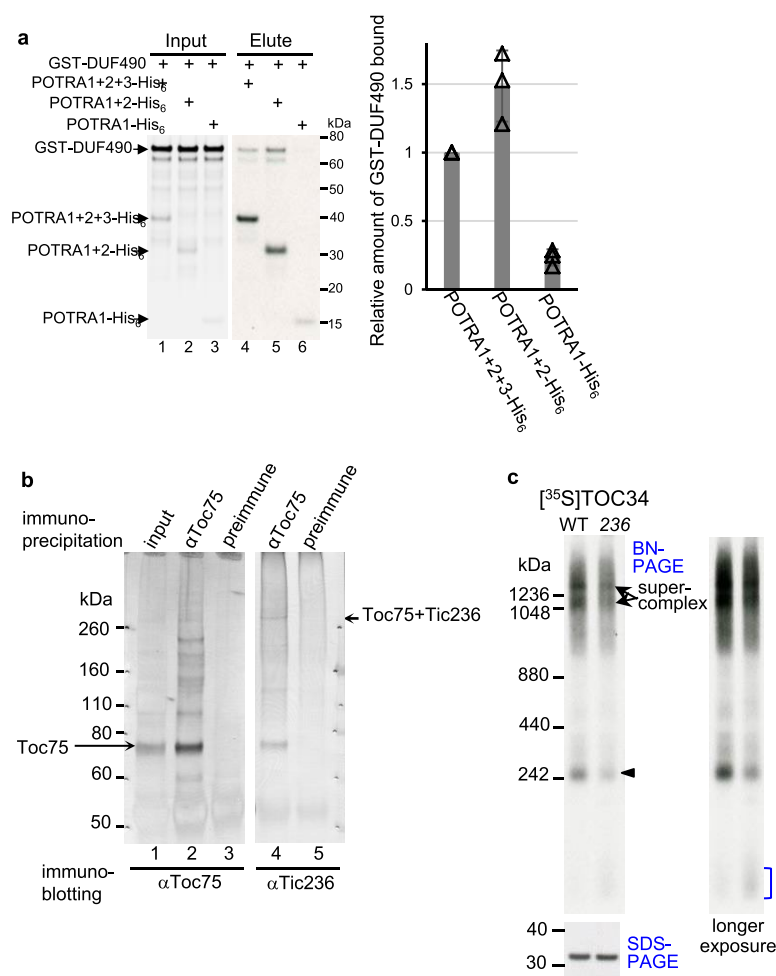


Extended Data Fig. 5 | The *tic236-2* and *tic236-3* mutants have reduced amounts of *TIC236* RNA and *TIC236* proteins. **a, Total RNA was isolated from *tic236-2* and *tic236-3* mutants and their corresponding wild types grown on MS plates for 14 days under 16-h light. Levels of *TIC236* RNA expression were analysed by quantitative RT-PCR. Levels of *UBQ10* RNA were analysed as normalization controls. Data are from two independent plant batches with three technical repeats for each batch, and are calculated by the Bio-Rad CFX Manager 3.1 software and shown as means \pm s.e. The original Cq (quantification cycles) for the six samples are: *Col* (*UBQ10*): 20.82, 20.42, 20.38, 19.33, 19.42 and 19.31; *Col* (*TIC236*): 31.35, 31.9, 32.22, 30.64, 30.57 and 30.49; *tic236-2* (*UBQ10*): 19.87, 19.86, 20.07, 18.64, 18.62 and 18.75; *tic236-2* (*TIC236*): 34.6, 34.2, 34.8, 33.24, 32.77 and 32.96; *No-0* (*UBQ10*): 19.21, 19.22, 19.27, 18.34, 18.36 and 18.37; *No-0* (*TIC236*): 30.09, 30.41, 30.31, 29.42, 29.74 and 29.34; *tic236-3* (*UBQ10*): 18.13, 17.96, 18.14, 17.63, 17.66 and 17.64; *tic236-3* (*TIC236*): 32.51, 32.86, 32.47, 31.97, 32.62 and 32.1. **b**, Total proteins were extracted from 20-day-old plants grown on MS plates under 16-h light, and analysed by SDS-PAGE (20 μ g of proteins per lane) and immunoblotting with the antibodies indicated at right. SPS, sucrose phosphate synthase; cpHSC70, chloroplast stromal HSC70. Data shown are representative of two independent experiments.**



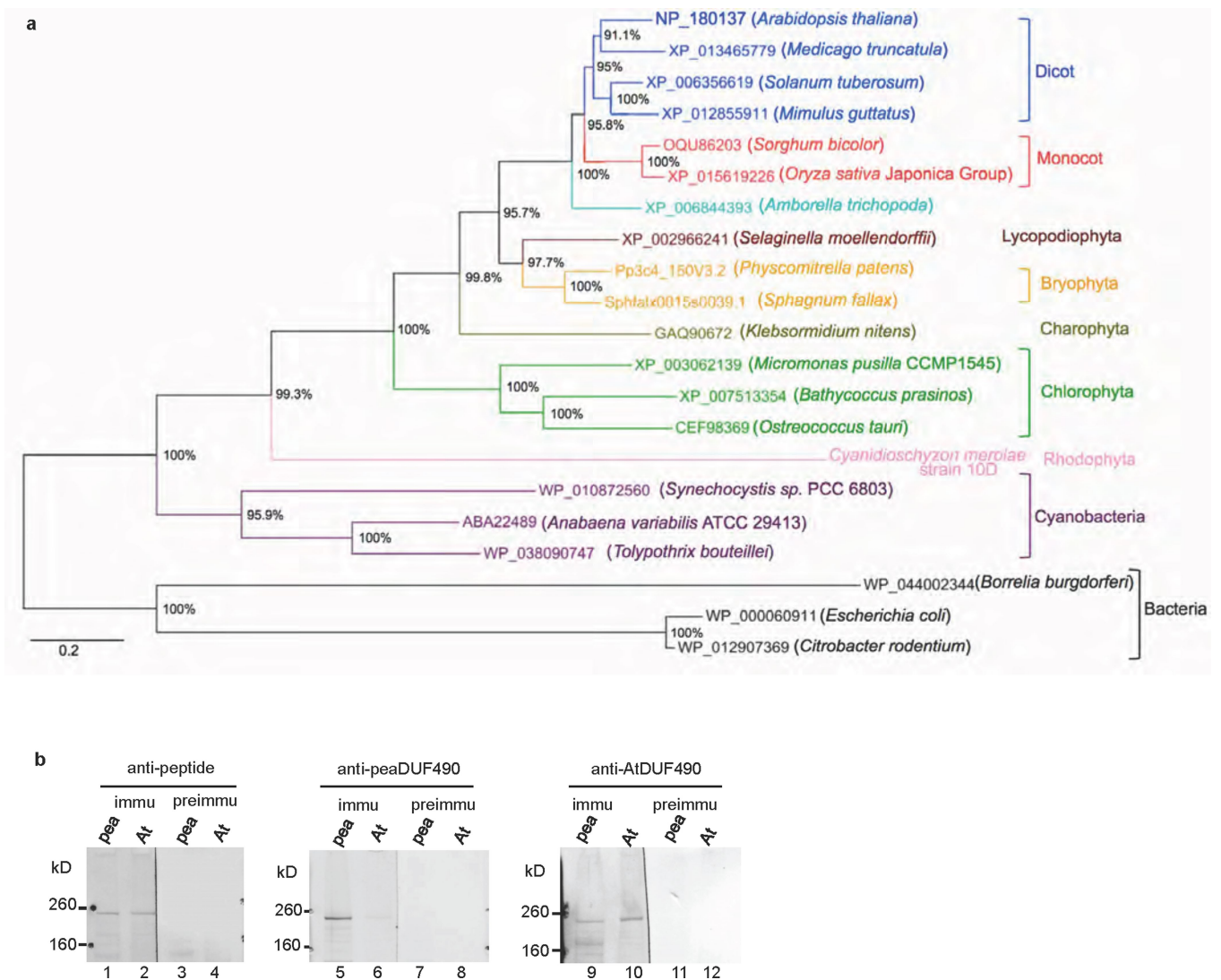
Extended Data Fig. 6 | *tic236*-mutant chloroplasts accumulate more preproteins on the chloroplast surface and have reduced rates of protein import into the stroma. **a**, $[^{35}\text{S}]$ prRBCS and $[^{35}\text{S}]$ prHSP93 were imported into chloroplasts isolated from 14-day-old *tic236-3* mutant and the corresponding wild-type (*No-0*) plants, under 1 mM ATP at room temperature for 10 min. After import, half of the chloroplasts were further treated with thermolysin. Re-isolated intact chloroplasts were analysed by SDS-PAGE and the gels were stained with Coomassie blue. Fluorographs are shown above respective images from the same gel. **b**, $[^{35}\text{S}]$ Met-labelled prTIC40 and prHSP93 were incubated with isolated wild-type (WT) and *tic236* (236) mutant chloroplasts under 1 mM ATP at room temperature for various amounts of time, as indicated above the gel. Chloroplasts were re-isolated and analysed by SDS-PAGE and the gels were stained with Coomassie blue. Fluorographs are shown above respective images from

the same gel. The precursor, intermediate and mature forms of prTIC40 are indicated by pr, i and m, respectively. Quantification of the amount of imported mature proteins (mature plus intermediate proteins for TIC40), corrected for loading by quantifying the amount of RBCL (for prHSP93) or CAB (for prTIC40) and normalized to the amount imported in the wild-type chloroplasts at 20 min, is shown at right. **c**, *tic236-2*-knockdown mutant plants have similar amounts of major translocon components to wild-type plants (*Col*). Total leaf proteins were extracted from 20-day-old plants and analysed by SDS-PAGE and immunoblotting with the antibodies labelled at right. '1x' represents 5 μg for the anti-TIC20 blot and 2.5 μg for other blots. Cytosolic SPS was also analysed as a control. Data shown as mean \pm s.d. of three independent experiments (**b**); representative of three (**c**) and two (**a**) independent experiments.



Extended Data Fig. 7 | TIC236 directly binds TOC75 and *Arabidopsis* *tic236*-mutant chloroplasts have reduced amounts of TOC-TIC supercomplexes. **a**, GST-DUF490 was incubated with a quarter amount of POTRA1-POTRA2-POTRA3-His₆ (POTRA1+2+3-His₆), POTRA1-POTRA2-His₆ (POTRA1+2-His₆) or POTRA1-His₆. Proteins pulled down by metal-affinity resin were analysed by SDS-PAGE and fluorography. Equal moles of POTRA proteins were loaded among the lanes. The amount of GST-DUF490 pulled down was quantified. **b**, Membranes from pea chloroplasts treated with 0.5 mM SMCC were solubilized by 1% LDS. The clarified supernatant (input) was immunoprecipitated with anti-TOC75 or the preimmune serum and protein A beads, analysed by SDS-PAGE and immunoblotting, and hybridized to anti-TOC75 or anti-TIC236 antibodies. Similar results were obtained for *Arabidopsis* chloroplasts (see Fig. 3b). **c**, We attempted to directly compare the amount of TOC-TIC supercomplexes in mutant and wild-type chloroplasts. However, although we could detect TOC-TIC supercomplex on one-dimensional BN-PAGE immunoblots using anti-TOC75 antibodies¹², blotting efficiencies from native gels were

variable and not quantitative. Consequently, we used imported [³⁵S] TOC34 to indirectly reflect the amounts of the translocon complexes. [³⁵S] TOC34 was imported into wild-type (WT) and *tic236-2* (236) mutant chloroplasts and analysed by BN-PAGE or SDS-PAGE. The two images in the BN-PAGE panel are from the same gel with different exposure times. In wild-type chloroplasts, TOC34 was detected in the 1.25- and 1-megadalton supercomplexes and in a complex about 242 kDa in size (arrowhead). In the *tic236*-mutant chloroplasts, the same complexes were detected but their amounts were reduced, and the amount of unassembled TOC34 migrating below 200 kDa (bracket) increased. Because the *tic236*-knockdown mutation does not affect the insertion of TOC34 into the outer membrane (Fig. 2e), these results suggest that a lower amount of the supercomplexes was present in the *tic236*-mutant chloroplasts. However, we cannot exclude the possibility that mutant chloroplasts were defective in the assembly of TOC34 into supercomplexes. The composition of the 242-kDa, TOC34-containing complex also remains to be determined. Data shown as mean \pm s.d. of three independent experiments (**a**); representative of two (**b**, **c**) independent experiments.



Extended Data Fig. 8 | Lineage distribution of TIC236 and specificity of the three anti-TIC236 antibody preparations. **a**, Phylogenetic relationships of sequences of the DUF490 of TIC236, from bacteria to higher plants. Bootstrap values from 1,000 replicates are indicated. The 0.2 scale shows substitution distance. **b**, Specificity of the three anti-TIC236 antibody preparations. Pea and *Arabidopsis* (At) total chloroplast proteins were analysed by SDS-PAGE (10 μ g per lane) and immunoblotting, and hybridized to the three antibody preparations against TIC236—a mouse

anti-serum against a synthetic peptide corresponding to residues 1957 to 1988 of *Arabidopsis* TIC236 (anti-peptide), a rabbit antiserum against the DUF490 domain of pea TIC236 (anti-peaDUF490), and a rabbit antiserum against the DUF490 domain of *Arabidopsis* TIC236 (anti-AtDUF490) (labelled 'immu')—as well as their corresponding preimmune sera (labelled 'preimmu'). Data shown are representative of at least two independent experiments.

Extended Data Table 1 | Forty-eight species and the accession numbers for sequences used for the MirrorTree analyses

Extended Data Table 1 Forty eight species and the accession numbers for sequences used for the MirrorTree analyses

Organism	Taxonomy ID	Database	Accession number (Tic236 ortholog)	Phytozome accession (Tic236 ortholog)	Accession number (Toc75 ortholog)	Phytozome accession (Toc75 ortholog)
{Cyanobacteria}						
<i>Anabaena variabilis</i> ATCC 29413	240292	Genbank	ABA22489		ABA19720	
<i>Cyanothece</i> sp. PCC 7425	395961	Genbank	WP_012630209		WP_012629853	
<i>Gloeocapsa</i> sp. PCC 7428	1173026	Genbank	AFZ29093		AFZ31056	
<i>Synechococcus</i> sp. UTEX 2973	1350461	Genbank	AJD57697		AJD58518	
<i>Tolypothrix bouleillei</i>	1246981	Genbank	WP_038090747		WP_038074595	
<i>Trichodesmium erythraeum</i> IMS101	203124	Genbank	ABG49815		ABG53623	
{Rhodophyta}						
<i>Cyanidioschyzon merolae</i> strain 10D	280699				BAM80320	
{Chlorophyta}						
<i>Auxenochlorella protothecoides</i>	3075	Genbank	XP_011402215		XP_011398083	
<i>Bathycoccus prasinos</i>	41875	Genbank	XP_007513354		XP_007509108.1	
<i>Chlorella variabilis</i>	554065	Genbank	XP_005847013		XP_005843553	
<i>Coccomyxa subellipsoidea</i>	574566	Both	XP_005645734	estExt_Genemark1.C_130201	XP_005649815	13722
<i>Micromonas pusilla</i> CCMP1545	564608	Genbank	XP_003062139		XP_003059853	
<i>Micromonas</i> sp. RCC299	296587	Both	XP_002499772	gw2.02.17.1	XP_002504526	102215
<i>Ostreococcus tauri</i>	70448	Genbank	CEF98369		XP_003074813	
{Charophyta}						
<i>Nitella mirabilis</i>	231897	Genbank, TSA database	JV739677		JV769324	
{Embryophyta}						
<i>Amborella trichopoda</i>	13333	Both	XP_006844393	evm_27.model.AmTr_v1.0_scaffold00142.59	XP_006846297	evm_27.model.AmTr_v1.0_scaffold00012.260
<i>Aquilegia coerulea</i>	218851	Pytozome		Aquca_001_00602.1		Aquca_003_00131.1
<i>Arabidopsis thaliana</i>	3702	Both	NP_180137	AT2G25660.1	NP_190258	AT3G46740.1
<i>Boechera stricta</i>	72658	Pytozome		Bostr.26326s0087.1		Bostr.18473s0284.1
<i>Brachypodium distachyon</i>	15368	Pytozome	XP_003569793	Bradi2g05017.1		Bradi1g66820.1
<i>Brassica rapa</i>	3711	Pytozome		Brara.D01549.1		Brara.F01795.1
<i>Capsella grandiflora</i>	264402	Pytozome		Cagra.8436s0002.1		Cagra.0448s0020.1
<i>Carica papaya</i>	3649	Pytozome		evm.model.supercontig_91.7		evm.model.supercontig_81.103
<i>Citrus sinensis</i>	2711	Pytozome		orange1.1g000108m		orange1.1g039285m
<i>Cucumis sativus</i>	3659	Genbank	XP_011652499		XP_004153150	
<i>Eucalyptus grandis</i>	71139	Genbank	KCW81498		XP_010043500	
<i>Kalanchoe laxiflora</i>	1670617	Pytozome		Kalax.0820s0008.1		Kalax.0329s0017.1
<i>Manihot esculenta</i>	3983	Pytozome		Manes.13G069700.1		Manes.01G169900.1
<i>Medicago truncatula</i>	3880	Both	XP_013465779	Medtr1g010300.1	XP_003606719	Medtr4g064780.1
<i>Mimulus guttatus</i>	4155	Both	XP_012855911	Migut.A00272.1	XP_012835500	Migut.D00882.1
<i>Musa acuminata</i>	4641	Genbank	XP_009417048		XP_009421330	
<i>Panicum hallii</i>	206008	Pytozome		Pahal.E04202		Pahal.H02553.1
<i>Phaseolus vulgaris</i>	3885	Pytozome		Phvul.001G055800.1		Phvul.005G146400.1
<i>Physcomitrella patens</i>	3218	Pytozome		Pp3c4_150V3.2		Pp3c6_10010V3.1
<i>Populus trichocarpa</i>	3694	Pytozome		Potri.018G034700.1		Potri.001G072800.1
<i>Prunus persica</i>	3760	Genbank	XP_007221927		XP_007208340	
<i>Ricinus communis</i>	3988	Pytozome		29794.m003487		29751.m001894
<i>Selaginella moellendorffii</i>	88036	Both	XP_002966241	439584	XP_002968046	270688
<i>Setaria viridis</i>	4556	Pytozome		Sevir.5G124500.1		Sevir.9G464600.1
<i>Solanum lycopersicum</i>	4081	Genbank	XP_010325153		XP_004241213	
<i>Solanum tuberosum</i>	4113	Genbank	XP_006356619		XP_006350787	
<i>Sorghum bicolor</i>	4558	Pytozome		Sobic.003G043200.1		Sobic.001G423300.1
<i>Sphagnum fallax</i>	53036	Pytozome		Sphfalx0015s0039.1		Sphfalx0149s0034.1
<i>Spirodela polyrrhiza</i>	29656	Pytozome		Spipo21G0013000		Spipo22G0038600
<i>Theobroma cacao</i>	3641	Both	XP_007013733	Thecc1EG038299t1	XP_007016346	Thecc1EG041787t1
<i>Vitis vinifera</i>	29760	Genbank	CB120936		XP_002280661	
<i>Zea mays</i>	4577	Both	DAA53164	GRMZM2G083374_T02	NP_001168264	GRMZM2G001918_T01
<i>Zostera marina</i>	29655	Both	KMZ69724	Zosma208g00170	KMZ66689	Zosma28g00470.1

Extended Data Table 2 | List of primers used in this study

Experiments	primers	Sequence (5'-3')	Cloning site	Target
atTic236 cDNA cloning	M4-Sp-F	ATCG <u>ACTAGT</u> ATGAGTTTGAGATTGCAAAACCC	<i>SpeI</i>	Arabidopsis Tic236 (At2g25660)
	M4-Xh-R	ATCG <u>CTCGAG</u> GTCTTGTGATGTAGCAGAGTATTC	<i>XhoI</i>	Arabidopsis Tic236 (At2g25660)
	Interlace PCR to generate sense mutation	M4-BamHI-F	BamHI	Arabidopsis Tic236 (At2g25660)
	M4-EvXhoI-R	GATCCTCGAGATGTGATATCACAATATCTCC	XhoI and EcoRV	Arabidopsis Tic236 (At2g25660)
	M4-dglI-F	TCTAAGGGAAGAGCTACxGGxGAXGTxCATCTATGTATGTCTAG		Arabidopsis Tic236 (At2g25660)
	M4-dglI-R	CTAGACATACATAGATGxACxTCxCCxGTAGCTCTTCCCTTAGA		Arabidopsis Tic236 (At2g25660)
<i>tic236</i> mutants identification				
<i>tic236-1</i>	SALK_0487700-RP	CATCGAGGCTAGAAATTGCAG		Arabidopsis Tic236 (At2g25660)
	LB-a1	TGGTTACGTAAGTGGGCCATCG		T-DNA left border
<i>tic236-2</i>	At2g25660-gDNA-F	ATATTTAGACTCCACCTGAACCAAC		Arabidopsis Tic236 (At2g25660)
	A-SEQ-1-R	GCCTATCTCTGCAGCTTTTCTTGC		5' UTR
	Syngenta-LB3	TAGCATCTGAATTTTATAACCAATCTCGATACAC		Arabidopsis Tic236 (At2g25660) (exon1)
<i>tic236-3</i>	Pst01662-primerA	TTGCACATTTTCCACAAA		T-DNA left border
	Pst0216-primerB	AAACGTGCTTGACTCGACCT		Arabidopsis Tic236 (At2g25660) (exon1)
	RIKEN-3-1a	GGTTCCCGTCCGATTTTCGACT		Arabidopsis Tic236 (At2g25660) 5' UTR
q-PCR	q-atM4-F	CAGGGAATGCTTGGATCTGT		Ds transposon
	q-atM4-R	AGTTCACATCCAGTGCGGTA		Arabidopsis Tic236 (At2g25660)
	UBQ10-F	TCCGGATCAGCAGAGGCTTA		Arabidopsis Tic236 (At2g25660)
	UBQ10-R	TCAGAACTCTCCACCTCAAG		UBQ10 (At4g05320)
fusion protein constructs				
GST-DUF490	M4-DUF490-EcoRI-F	ATGCGAATTCCTACTGGAAAGCAGGGTAAGTAG	<i>EcoRI</i>	Arabidopsis Tic236 (At2g25660)
	M4-DUF490-R	ATGCACTAGTTTGTGCTTGTGATGTAGCAGAGTATTC		DUF490 domain
GST-DUF490DC	M4-DUF490DC-R	ATGCACTAGTTTAAAGATTGCAGAAGGACACGCAAG	<i>SpeI</i>	Arabidopsis Tic236 (At2g25660)
				DUF490 domain
His ₆ -atDUF490	At490-XhoI-F	CTAGCTCGAGCACCTCACTGGAAGCAGGGTAAGTAG	<i>XhoI</i>	Arabidopsis Tic236 (At2g25660)
	At490-PstI-R	GACTCTGCAGTTAGTCTTGTGATGTAGCAGAGTATTC		DUF490 domain
His ₆ -peaDUF490	Pea490-XhoI-F	CTAGCTCGAGCACATCACTAGAGAGTAGGCTAAGCAG	<i>XhoI</i>	Arabidopsis Tic236 (At2g25660)
	Pea490-PstI-R	GACTCTGCAGTCAATCCTGAGATGTGGCAGAATATTC		DUF490 domain
POTRA1+2-His6	75DP3-F	GTGAAGTTGTGGAAGGTGATCTAGAACAGAAGTCAGCTG		Arabidopsis Tic236 (At2g25660)
	75DP3-R	CAGCTGACTTCTGTTCTAGATCACCTTCCACAACCTTCAC		POTRA domains
POTRA1-His6	75DP23-F	CTCGTTTGCTGAGAGTACACTAGAACAGAAGTCAGCTG		Arabidopsis Tic236 (At2g25660)
	75DP23-R	CAGCTGACTTCTGTTCTAGTGTACTCTCAGCAAACGAG		POTRA domains
Transit peptide estimation				
Site-directed mutagenesis	M4L228STOP-F	GATTTTACATGGTTAGGGATACCTTAATCTGACACTACTTTGCCAAGCC		Arabidopsis Tic236 (At2g25660)
	M4L228STOP-R	GGCTTGGCAAAGTAGTGTCAGATTAAGGTATCCCTAACCATGTAAAATC		cDNA, residue 218, K ot Stop
	F-M41-227del	GTGACACTATAGAATCGAGATAAGTTTGAGATTGCAAAAC		Arabidopsis Tic236 (At2g25660)
	R-M41-227del	GTTTTGCAATCTCAAACCTTATCTCGAGTTCTATAGTGTCAC		cDNA, residue 1, M ot I
	F-M4R30M	GCAGAGAGAAGAGAATCAATGTAGCTATGAGGGCGTTTCGTAG		Arabidopsis Tic236 (At2g25660)
	R-M4R30M	CTACGAAACGCCCTCATAGCTACATTGATTCTTCTCTCTGCG		cDNA, residue 30, R ot M
	F-M4S35M	GCTAGAAGGGCGTTTCGTATGAAGCGTATATATTCG		Arabidopsis Tic236 (At2g25660)
	R-M4S35M	CGAATATATACGCTTCATACGAAACGCCCTTCTAGC		cDNA, residue 30, R ot M
	F-M4I38M	CGTTTCGTAGCAAGCGTATGTATTCGGAGAAGAAACAGA		Arabidopsis Tic236 (At2g25660)
	R-M4I38M	TCTGTTTCTTCTCGAATACATACGCTTGCTACGAAACG		cDNA, residue 35, S ot M
	F-M4D46M	CGGAGAAGAAACAGAAATATGTGGTTAGCTAAAGTTGCC		Arabidopsis Tic236 (At2g25660)
	R-M4D46M	GGCAACTTTAGCTAACACATATTCTGTTTCTTCTCCG		cDNA, residue 35, S ot M
	F-M4S55M	GGTTAGCTAAAGTTGCCAAATTTATGCAATTTTGTTGGG		Arabidopsis Tic236 (At2g25660)
	R-M4S55M	CCCACAAAATTGCATAAATTTGGCAACTTTAGCTAAC		cDNA, residue 46, D ot M
				Arabidopsis Tic236 (At2g25660)
				cDNA, residue 46, D ot M
				Arabidopsis Tic236 (At2g25660)
				cDNA, residue 55, S ot M
				Arabidopsis Tic236 (At2g25660)
				cDNA, residue 55, S ot M

CDK12 regulates DNA repair genes by suppressing intronic polyadenylation

Sara J. Dubbury^{1,2,4}, Paul L. Boutz^{1,3,4} & Phillip A. Sharp^{1,2*}

Mutations that attenuate homologous recombination (HR)-mediated repair promote tumorigenesis and sensitize cells to chemotherapeutics that cause replication fork collapse, a phenotype known as ‘BRCAness’¹. BRCAness tumours arise from loss-of-function mutations in 22 genes¹. Of these genes, all but one (CDK12) function directly in the HR repair pathway¹. CDK12 phosphorylates serine 2 of the RNA polymerase II C-terminal domain heptapeptide repeat^{2–7}, a modification that regulates transcription elongation, splicing, and cleavage and polyadenylation^{8,9}. Genome-wide expression studies suggest that depletion of CDK12 abrogates the expression of several HR genes relatively specifically, thereby blunting HR repair^{3–7,10,11}. This observation suggests that the mutational status of CDK12 may predict sensitivity to targeted treatments against BRCAness, such as PARP1 inhibitors, and that CDK12 inhibitors may induce sensitization of HR-competent tumours to these treatments^{6,7,10,11}. Despite growing clinical interest, the mechanism by which CDK12 regulates HR genes remains unknown. Here we show that CDK12 globally suppresses intronic polyadenylation events in mouse embryonic stem cells, enabling the production of full-length gene products. Many HR genes harbour more intronic polyadenylation sites than other expressed genes, and these sites are particularly sensitive to loss of CDK12. The cumulative effect of these sites accounts for the enhanced sensitivity of HR gene expression to CDK12 loss, and we find that this mechanism is conserved in human tumours that contain loss-of-function CDK12 mutations. This work clarifies the function of CDK12 and underscores its potential both as a chemotherapeutic target and as a tumour biomarker.

CDK12 regulates HR gene expression by an unknown mechanism. Mouse embryonic stem (mES) cells are primarily in S-phase of the cell cycle and fail to activate a G1/S checkpoint after DNA damage, making them reliant on replication-coupled HR repair and sensitive to HR defects^{12–14}. We sought to dissect the molecular function of CDK12 by generating *Cdk12* genetic knockouts (*Cdk12Δ*) in mES cells that express a complementing, doxycycline (Dox)-inducible *Cdk12* transgene under continuous Dox treatment (Extended Data Fig. 1a, b). Upon withdrawal of Dox, CDK12 was depleted after 24 h and undetectable after 48 h (Fig. 1a, Extended Data Fig. 1c). Loss of CDK12 yielded a progressive cell viability defect after 72 h of Dox depletion, which was reversible upon CDK12 re-expression (Fig. 1b, Extended Data Fig. 1d). Notably, the initial 48 h of CDK12 depletion had minimal consequences on viability, providing a window in which to probe CDK12 function.

The cell viability defect observed upon CDK12 loss could be due to decreased proliferation and/or increased cell death. Cell cycle profiling upon CDK12 depletion revealed decreased nucleotide incorporation during S-phase and a shift in the proportion of cells from S-phase to G1, which was reversed upon re-expression of CDK12 (Fig. 1c, Extended Data Fig. 1e). In addition, the percentage of cells undergoing apoptosis increased upon CDK12 loss (Fig. 1d, Extended Data Fig. 1f). Failure to repair DNA damage during S-phase causes replication fork stalling and impaired DNA replication¹⁵, which is consistent with the decreased

nucleotide incorporation during S-phase observed upon CDK12 depletion. Persistent DNA damage forces mES cells to differentiate or initiate apoptosis^{16,17}. The accumulation of cells in G1 after CDK12 loss is consistent with differentiating cells that have longer G1 phases and competent G1/S checkpoints¹⁸, and the increase in apoptosis is consistent with programmed cell death in response to unrepaired DNA damage. Indeed, withdrawal of Dox for 48 h resulted in the accumulation of DNA double-strand breaks (Fig. 1e, Extended Data Fig. 1g). Furthermore, both total p53 and Ser15-phosphorylated (activated) p53¹⁹ were upregulated upon CDK12 loss (Fig. 1f, Extended Data Fig. 1h). Thus, CDK12 ablation results in phenotypes consistent with defective HR repair in mES cells.

To address the molecular consequences of CDK12 loss, we sequenced RNA after 24 and 48 h of CDK12 depletion. One hundred and forty genes (after 24 h) and 814 genes (after 48 h) showed significant changes in total gene expression (posterior probability of differential expression (PPDE) > 0.95) (Extended Data Fig. 2a, Supplementary Table 1). Corroborating the p53 activation observed upon CDK12 loss, approximately 33% of these genes were validated p53 targets that changed in the expected direction²⁰ (Extended Data Fig. 2b). Consistent with induction of mES cell differentiation by p53 activation, there was also an enrichment of genes (12%) harbouring bivalent chromatin modifications (trimethylation of lysine 4 or lysine 27 on histone H3 (H3K4me3 and H3K27me3, respectively)), a marker of early differentiation genes, at their promoters²¹. These two gene signatures accounted for approximately 70% of the genes whose expression increased and 20% of the genes whose expression decreased upon CDK12 depletion. When we excluded these genes as likely secondary effects, we found that CDK12 depletion after 48 h modestly affected the total expression of only 428 (3%) of expressed genes, and most of those genes showed decreased expression upon CDK12 loss (Extended Data Fig. 2b).

In addition to gene expression changes at the total transcript level, alterations in isoform usage also affect functional gene output. Although CDK12 has been implicated in alternative splicing²², we observed few splicing changes (Extended Data Fig. 2c). We next examined alternative cleavage and polyadenylation. In addition to the polyadenylation site located after the 3′-most exon of a gene (the distal polyadenylation site), intronic polyadenylation sites (IPAs) occur throughout introns. Usage of IPAs produces truncated mRNA isoforms that vary in coding potential, stability, translational efficiency, and localization^{23,24}. Using 3′-end sequencing data from mES cells²⁵ to define high-confidence IPA and distal isoforms genome-wide, we found that CDK12-depleted cells showed global increases in IPA events at the expense of distal sites (Fig. 2a, b); this finding was validated by isoform-specific quantitative PCR following reverse transcription (RT-qPCR; Extended Data Fig. 2d–g). Among 33,115 IPAs identified in 13,594 expressed genes²⁵, 2,009 individual IPA isoforms (about 6.4% of identified IPA isoforms) were significantly ($P_{\text{adj}} < 0.05$) differentially expressed upon CDK12 loss (Extended Data Fig. 2h, Supplementary Tables 2, 3). The vast majority of these IPA isoforms (1,824, 91%) increased upon CDK12 depletion.

¹Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA, USA. ²Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA.

³Present address: Department of Biochemistry and Biophysics, University of Rochester School of Medicine and Dentistry, Rochester, NY, USA. ⁴These authors contributed equally: Sara J. Dubbury, Paul L. Boutz. *e-mail: sharp@mit.edu

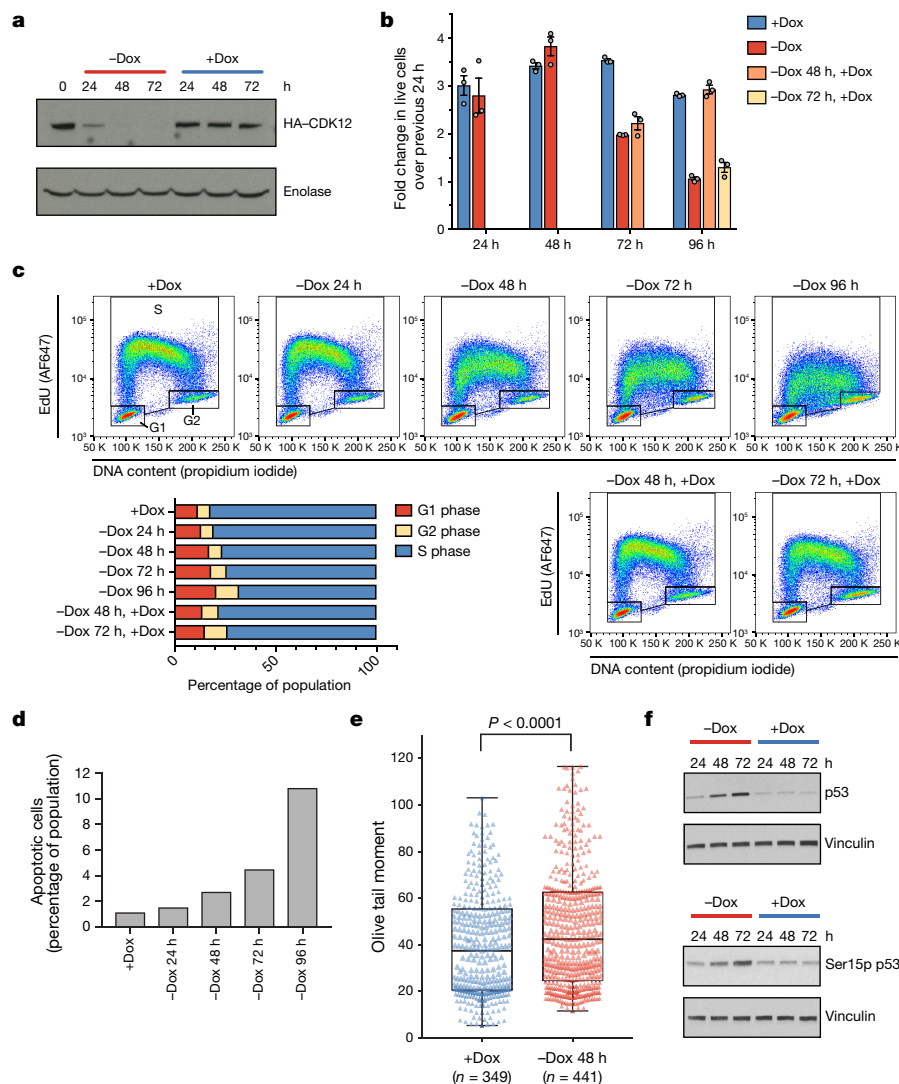


Fig. 1 | CDK12 depletion causes attenuated DNA damage repair in mES cells. **a–f**, Phenotypic data from one *Cdk12* Δ clone. **a**, Representative immunoblot for CDK12 (HA-CDK12) after Dox withdrawal. **b**, Fold change in live cells over previous 24 h. Bars show mean fold change (\pm s.e.m., $n = 3$ biological replicates) for cells grown in Dox continuously (blue), off Dox starting at time 0 (red), or off Dox beginning at time 0 and reintroduced to Dox after 48 h (orange) or 72 h (yellow) for the remainder of the experiment. **c**, Fluorescence-activated cell sorting (FACS) cell cycle profiling of one representative biological replicate for the same conditions

To quantify the expression of isoforms that result from the use of distal polyadenylation sites, we measured the expression of the distal-most exon as normalized to the rest of the transcript (Fig. 2a). In contrast to IPA isoforms, the majority (1,848, 75%) of significantly ($P_{\text{adj}} < 0.05$) changing distal isoforms decreased upon CDK12 loss (Fig. 2a, Extended Data Fig. 2h, Supplementary Table 4). In a subset of these genes (571), we could detect a corresponding, statistically significant increase in at least one IPA isoform (Extended Data Fig. 2i). The majority (56%) of the remaining genes that showed a decrease in expression of the distal isoform contained at least one IPA that increased in usage, even though they did not reach statistical significance. Whereas individual IPAs may not reach statistical significance owing to the presence of multiple IPA events within the same gene, the decrease in distal polyadenylation site usage represents the cumulative loss from each upstream IPA. Therefore, we consider these 1,848 genes with significantly decreasing distal exons also to be altered by CDK12-dependent IPA usage.

An intron may contain multiple IPAs, and a gene may contain multiple introns with IPAs. Collapsing the data to single genes indicated

as in **b**, quantified in bar plot. **d**, Quantification of apoptotic cells upon CDK12 loss for one representative experiment. **e**, Comet assay for DNA double-strand breaks in *Cdk12* Δ cells after 48 h of Dox withdrawal. Olive tail moment: the product of the percentage of total DNA in the tail and the distance between the intensity-weighted centroids of the comet head and tail. Box plots: median value with 25th and 75th quartiles, whiskers show minimum to maximum. P value based on one-sided Mann–Whitney U test. **f**, Immunoblot of total and Ser15 phosphorylated (Ser15p) p53 upon CDK12 loss.

that 2,948 genes (about 22% of expressed genes) had at least one significantly increasing IPA isoform, a significantly decreasing distal isoform, or both (Fig. 2c). We focused on this set of genes for the later mechanistic studies; however, notably, for all genes with an IPA, as a population the IPA isoform usage increased and distal isoform usage decreased whether or not the individual sites reached statistical significance (Fig. 2d). Therefore, we conclude that the primary role of CDK12 is to suppress IPAs genome-wide and to promote expression of distal (full-length) isoforms.

As cleavage and polyadenylation occur co-transcriptionally and CDK12 phosphorylates the RNA polymerase II C-terminal domain (RNAPII CTD), we investigated whether CDK12 depletion resulted in changes to RNAPII or its Ser2 phosphorylation (Ser2p) status that might explain the increased IPA site usage. Using chromatin immunoprecipitation sequencing (ChIP), we mapped RNAPII and Ser2p RNAPII density genome-wide using two independent antibodies per target in biological duplicate. ChIP profiles from the two independent antibodies were highly similar (Extended Data Fig. 3). Therefore,

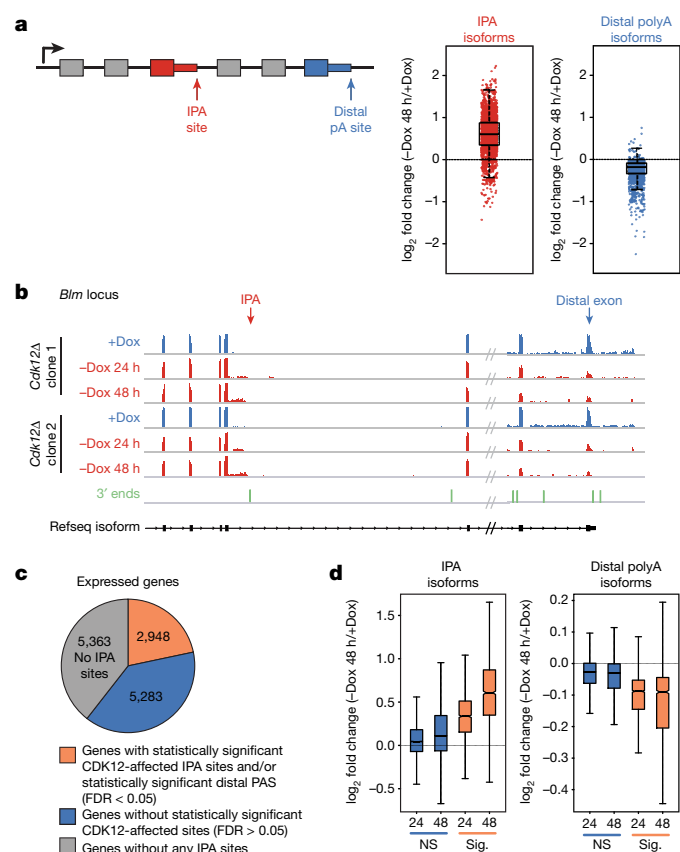


Fig. 2 | CDK12 loss increases IPA and decreases distal polyadenylation. **a**, Left, schematic showing an IPA and a distal polyadenylation site. Right, \log_2 fold change in normalized read density ($-\text{Dox } 48 \text{ h}/+\text{Dox}$) for IPA isoforms (red) and distal polyadenylation isoforms (blue) that reached statistical significance ($P_{\text{adj}} < 0.05$). **b**, RNA sequencing (RNA-seq) read density across the 3' end of *Blm* at one IPA site and at the distal exon; in Dox (blue, $n = 2$ biological replicates per clone) or after Dox withdrawal for 24 or 48 h (red, $n = 2$ biological replicates per time point and clone). 3' end sequencing read density below in green. **c**, Expressed genes with at least one significantly changing IPA and/or distal isoform (orange), with at least one IPA isoform with no significant change in IPA or distal isoforms (blue), or without any identified IPA sites (grey). **d**, \log_2 fold changes of all IPA sites (left) and all terminal sites (right) in expressed genes that changed significantly (sig.) upon Dox depletion for 24 or 48 h (orange) or that did not change significantly (NS) upon Dox depletion after 24 or 48 h (blue). **a**, **d**, FDR-adjusted P value determined by the DEXSeq package in R; $n = 4$ biological replicates for each condition. Box plots: median value with 25th and 75th quartiles, whiskers show $1.5 \times$ interquartile range.

the data from both antibodies and biological replicates were aggregated throughout the metagene analyses (Extended Data Fig. 4a). Furthermore, we developed a statistical framework (Extended Data Fig. 4b–d) for reliably measuring differences in ChIP read density.

To analyse the effects on RNAPII elongation, we plotted metagene profiles of RNAPII density from the transcriptional start site (TSS) to the distal polyadenylation site (Distal polyA) (Fig. 3a). Because RNAPII density correlates with gene expression levels and correlates inversely with gene length, we removed the shortest and longest length quartiles and focused on the middle two quartiles of genes (Extended Data Fig. 5a, b). Loss of CDK12 resulted in decreased RNAPII density at the 5' ends of genes, transitioning to increased RNAPII density towards the 3' ends (Fig. 3b, Extended Data Fig. 5b, c). This pattern was not specific to genes with statistically significant CDK12-sensitive IPA or distal polyA isoforms and was observed in length- and expression-matched control gene sets (Extended Data Fig. 6).

To determine whether decreased RNAPII density at the 5' ends of genes could be due to a reduction in RNAPII entering elongation, we aligned metagene profiles on the first stable nucleosome downstream

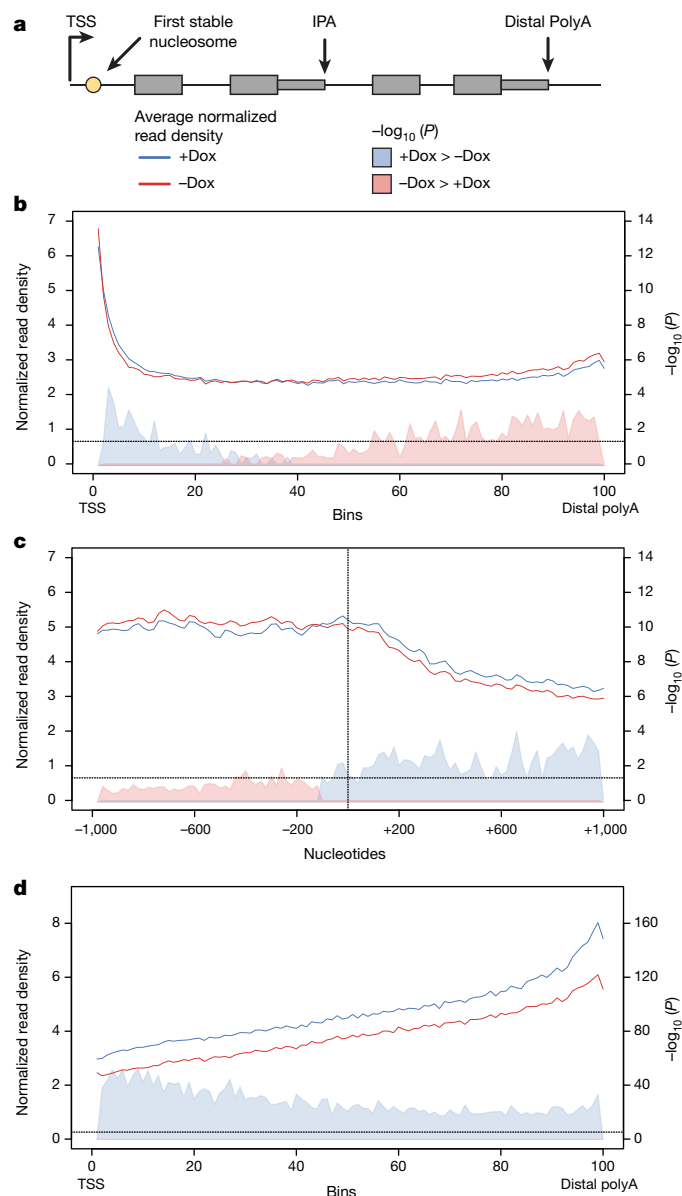


Fig. 3 | CDK12 loss results in altered RNAPII elongation dynamics and decreased RNAPII-CTD Ser2 phosphorylation. **a**, Schematic of gene elements (top) and key to metagene plots (bottom). **b**, Metagene profile of total RNAPII density from the TSS to the distal polyadenylation site. **c**, Total RNAPII metagene density 1 kb upstream and downstream of the first stable nucleosome dyad (dashed vertical line). **d**, RNAPII CTD Ser2p metagene density. Panels **b**–**d** include genes with significantly changing IPA or distal isoforms; solid lines indicate normalized read density with (blue, $n = 4$ independent ChIPs) or without (red, $n = 4$ independent ChIPs) CDK12; shaded areas indicate $-\log_{10}$ (bin-wise P value, Kolmogorov–Smirnov one-sided test) of the difference in read density (blue indicates that CDK12⁺ signal is greater, pink indicates that CDK12[−] signal is greater). Horizontal dashed line: $P = 0.05$. Shortest and longest gene length quartiles are excluded in **b**, **d** (Extended Data Fig. 5).

of the promoter, a barrier associated with RNAPII entering productive elongation²⁶ (Fig. 3a). Upon CDK12 loss, RNAPII density increased upstream and decreased downstream of the first stable nucleosome, accounting for the decrease in RNAPII density at the 5' ends of genes and indicating that less RNAPII entered productive elongation in these cells than in CDK12-expressing cells (Fig. 3c, Extended Data Fig. 7). Therefore, the increased RNAPII density towards the 3' ends of genes is most parsimoniously explained by altered elongation dynamics that cause progressive RNAPII accumulation across gene bodies upon CDK12 depletion. Consistent with its role as a RNAPII Ser2 kinase,

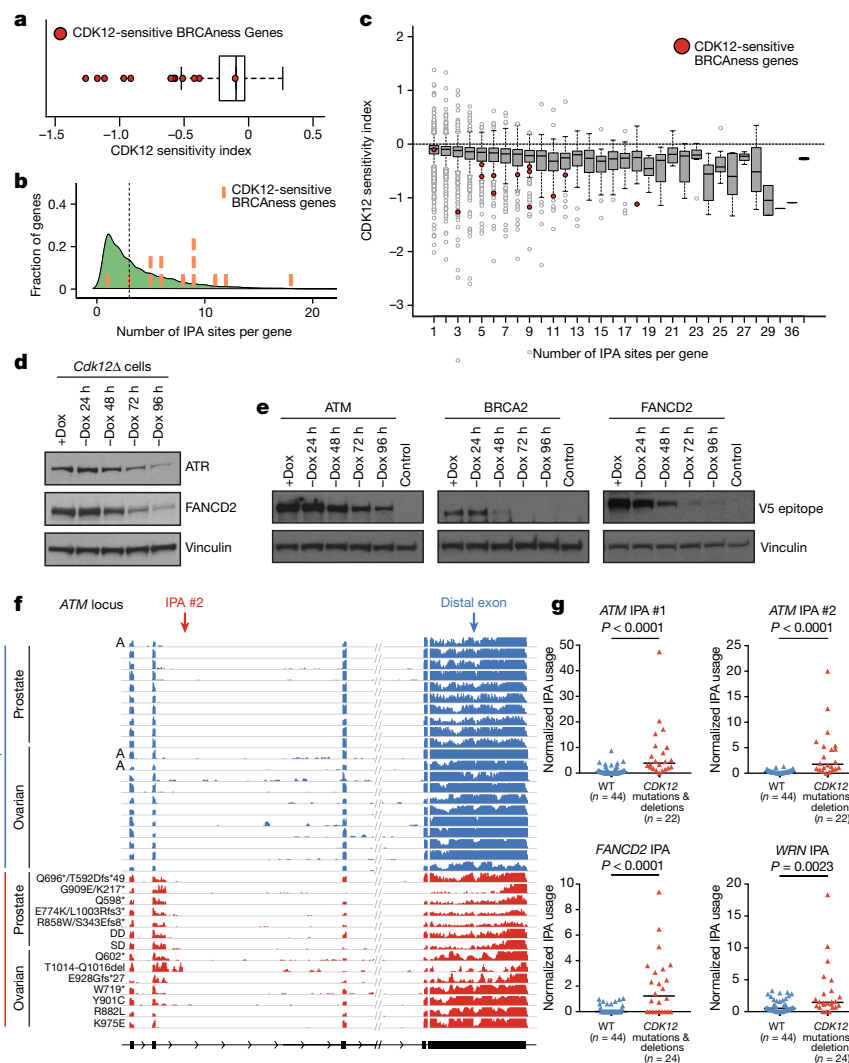


Fig. 4 | HR genes are highly responsive to CDK12 loss and human tumours with CDK12 LOF upregulate IPAs. **a**, Distribution of the CDK12 sensitivity index for all expressed genes with at least one IPA site. Red circles, CDK12-sensitive HR genes. Boxplots in **a**, **c**: median with 25th and 75th quartiles; whiskers show $1.5 \times$ interquartile range. **a**, **c**, $n = 4$ biological replicates per condition. **b**, Kernel density plot showing distribution of IPA sites per gene in expressed genes with at least one IPA site. CDK12-sensitive HR genes superimposed as orange bars (left to right: *Bap1*, *Atr*, *Fanc1*, *Wrn*, *Brca1*, *Brca2*, *Fancm*, *Brip1*, *Fancd2*, *Fanci*, *Blm*, *Fanca*, *Atm*). **c**, CDK12 sensitivity index for expressed genes grouped by number of IPA sites per gene. Red circles, CDK12-sensitive HR genes. **d**, Immunoblots of ATR and FANCD2 (endogenous antibodies) in *Cdk12*Δ cells after Dox removal. **e**, Immunoblots for representative clone of each cell line endogenously V5 epitope-tagged at *Atm*, *Brca2*,

and *Fancd2* in *Cdk12*Δ cells after Dox removal. Lysate from untagged *Cdk12*Δ cells (+Dox) is control. **f**, RNA-seq read density in *ATM* from TCGA (the cancer genome atlas) tumours with the indicated mutational status. Tumours shown in blue are wild-type for *CDK12* and diploid unless marked as amplified (A). Tumours shown in red carry missense putative driver mutations, truncating mutations, or shallow (SD) or deep (DD) *CDK12* gene deletions. All ovarian tumours with *CDK12* mutations (except R882L) also carry shallow deletions in *CDK12*. **g**, Quantification of IPA usage in *ATM* (2 different IPAs), *FANCD2* and *WRN* in TCGA tumours. Tumours with wild-type or amplified *CDK12* are shown in blue (WT), those with *CDK12* deletions, missense mutations, or truncating mutations in red (Mut). Black bars, medians. Sample size indicated below. *P* values from one-sided Mann–Whitney *U* test.

Ser2p RNAPII decreased across the entire gene body upon CDK12 loss (Fig. 3d, Extended Data Fig. 8). Therefore, our data suggest that CDK12-mediated phosphorylation of the RNAPII CTD is associated with a positive effect on transcription elongation dynamics across all or most genes. How this activity could alter IPA site usage is discussed in Extended Data Fig. 9.

We investigated whether enhanced IPA usage upon CDK12 depletion could account for the functional loss of HR that has been previously reported and that is reflected in our phenotypic data. Indeed, there is an enrichment for the BRCAness genes among those genes that show statistically significant increases in IPA usage or decreases in distal polyadenylation usage upon CDK12 loss (13 genes out of 22 BRCAness genes¹, $P = 1.59 \times 10^{-4}$, Fishers exact test). Furthermore, the distal polyadenylation isoforms of these 13 critical HR genes are more profoundly decreased as a group by CDK12 loss than the distal isoforms

of other genes (Fig. 4a). The enhanced sensitivity of HR genes to IPA activation is explained by two observations. First, the CDK12-sensitive HR genes are enriched in the frequency of IPAs per gene (Fig. 4b). In genes with multiple IPAs, the negative effect of terminating at each individual IPA on the amount of full-length isoform is cumulative; consequently, we observed a strong correlation between the number of IPAs per gene and the effect of CDK12 depletion on the production of the full-length isoform for that gene (Fig. 4c). Second, compared to expressed genes with the same number of IPAs, the HR genes showed increased sensitivity to CDK12 loss (Fig. 4a, c). Our data suggest that the cumulative effect of multiple, high-sensitivity IPAs in HR genes accounts for the downregulation of their full-length isoforms. These isoform changes substantially decreased the expression of full-length protein (Fig. 4d, e). As CDK12 activity maintains the full-length expression of more than half of the identified BRCAness genes, we propose

that the combined effect of strong downregulation of multiple gene products within the same functional pathway causes the HR-deficient phenotypes observed upon CDK12 loss.

Predicted and validated⁵ CDK12 loss-of-function (LOF) point mutations and deletions have been recurrently identified in prostate^{27,28} and ovarian^{29,30} tumours. RNA sequencing data from the tumours of patients with ovarian serous adenocarcinoma²⁹ and prostate adenocarcinoma²⁷ showed that putative CDK12 LOF mutations, but not oncogenic mutations in other BRCAness genes, increased IPA usage within key BRCAness genes, including *ATM*, *WRN*, and *FANCD2*, compared to tumours that were wild type for CDK12 (Fig. 4f, g; Extended Data Fig. 10a, b). Notably, the tumour harbouring the CDK12(K975E) missense mutation, which was previously validated to have a minimal effect on CDK12 activity⁵, did not show increased *ATM* IPA expression (Fig. 4f).

We treated prostate adenocarcinoma and ovarian carcinoma cell lines with the CDK12/CDK13 inhibitor THZ531⁴ to validate the increased IPA usage observed in CDK12 mutant tumours; as expected, IPA usage increased (Extended Data Fig. 10c–e). These data suggest that the role of CDK12 in suppressing IPA usage is conserved in humans and that tumours harbouring CDK12 loss-of-function mutations upregulate IPAs, abrogating functional HR. Differential IPA usage may therefore serve as a biomarker for functional CDK12 (and thus HR) loss in tumours that harbour uncharacterized CDK12 mutations and could potentially be used to identify patients who would respond to targeted treatments against BRCAness phenotypes, such as PARP1 inhibitors.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0758-y>.

Received: 26 December 2017; Accepted: 11 October 2018;

Published online 28 November 2018.

- Lord, C. J. & Ashworth, A. BRCAness revisited. *Nat. Rev. Cancer* **16**, 110–120 (2016).
- Bartkowiak, B. et al. CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1. *Genes Dev.* **24**, 2303–2316 (2010).
- Blazek, D. et al. The cyclin K/CDK12 complex maintains genomic stability via regulation of expression of DNA damage response genes. *Genes Dev.* **25**, 2158–2172 (2011).
- Zhang, T. et al. Covalent targeting of remote cysteine residues to develop CDK12 and CDK13 inhibitors. *Nat. Chem. Biol.* **12**, 876–884 (2016).
- Ekumi, K. M. et al. Ovarian carcinoma CDK12 mutations misregulate expression of DNA repair genes via deficient formation and function of the CDK12/CycK complex. *Nucleic Acids Res.* **43**, 2575–2589 (2015).
- Johnson, S. F. et al. CDK12 inhibition reverses de novo and acquired PARP inhibitor resistance in BRCA wild-type and mutated models of triple-negative breast cancer. *Cell Reports* **17**, 2367–2381 (2016).
- Iniguez, A. B. et al. EWS/FLI confers tumor cell synthetic lethality to CDK12 inhibition in Ewing sarcoma. *Cancer Cell* **33**, 202–216.e6 (2018).
- Hsin, J. P. & Manley, J. L. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes Dev.* **26**, 2119–2137 (2012).
- Davidson, L., Muniz, L. & West, S. 3' end formation of pre-mRNA and phosphorylation of Ser2 on the RNA polymerase II CTD are reciprocally coupled in human cells. *Genes Dev.* **28**, 342–356 (2014).
- Bajrami, I. et al. Genome-wide profiling of genetic synthetic lethality identifies CDK12 as a novel determinant of PARP1/2 inhibitor sensitivity. *Cancer Res.* **74**, 287–297 (2014).
- Joshi, P. M., Sutor, S. L., Huntoon, C. J. & Karnitz, L. M. Ovarian cancer-associated mutations disable catalytic activity of CDK12, a kinase that promotes homologous recombination repair and resistance to cisplatin and poly(ADP-ribose) polymerase inhibitors. *J. Biol. Chem.* **289**, 9247–9253 (2014).
- Hong, Y. & Stambrook, P. J. Restoration of an absent G1 arrest and protection from apoptosis in embryonic stem cells after ionizing radiation. *Proc. Natl Acad. Sci. USA* **101**, 14443–14448 (2004).
- Aladjem, M. I. et al. ES cells do not activate p53-dependent stress responses and undergo p53-independent apoptosis in response to DNA damage. *Curr. Biol.* **8**, 145–155 (1998).

- Tichy, E. D. et al. Mouse embryonic stem cells, but not somatic cells, predominantly use homologous recombination to repair double-strand DNA breaks. *Stem Cells Dev.* **19**, 1699–1711 (2010).
- Zeman, M. K. & Cimprich, K. A. Causes and consequences of replication stress. *Nat. Cell Biol.* **16**, 2–9 (2014).
- Lin, T. et al. p53 induces differentiation of mouse embryonic stem cells by suppressing Nanog expression. *Nat. Cell Biol.* **7**, 165–171 (2005).
- Liu, J. C. et al. High mitochondrial priming sensitizes hESCs to DNA-damage-induced apoptosis. *Cell Stem Cell* **13**, 483–491 (2013).
- van der Laan, S., Tsanov, N., Crozet, C. & Maiorano, D. High Dub3 expression in mouse ESCs couples the G1/S checkpoint to pluripotency. *Mol. Cell* **52**, 366–379 (2013).
- Shieh, S. Y., Ikeda, M., Taya, Y. & Prives, C. DNA damage-induced phosphorylation of p53 alleviates inhibition by MDM2. *Cell* **91**, 325–334 (1997).
- Lee, K.-H. et al. A genome-wide study identifies the Wnt signaling pathway as a major target of p53 in murine embryonic stem cells. *Proc. Natl Acad. Sci. USA* **107**, 69–74 (2010).
- Ku, M. et al. Genome-wide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet.* **4**, e1000242 (2008).
- Tien, J. F. et al. CDK12 regulates alternative last exon mRNA splicing and promotes breast cancer cell invasion. *Nucleic Acids Res.* **45**, 6698–6716 (2017).
- Tian, B. & Manley, J. L. Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.* **18**, 18–30 (2017).
- Tian, B., Pan, Z. & Lee, J. Y. Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Res.* **17**, 156–165 (2007).
- Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B. & Sharp, P. A. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**, 360–363 (2013).
- Weber, C. M., Ramachandran, S. & Henikoff, S. Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. *Mol. Cell* **53**, 819–830 (2014).
- Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer. *Cell* **163**, 1011–1025 (2015).
- Wu, Y. M. et al. Inactivation of CDK12 delineates a distinct immunogenic class of advanced prostate cancer. *Cell* **173**, 1770–1782.e14 (2018).
- Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
- Menghi, F. et al. The tandem duplicator phenotype is a prevalent genome-wide cancer configuration driven by distinct gene mutations. *Cancer Cell* **34**, 197–210.e5 (2018).

Acknowledgements We thank the Sharp laboratory, J. Arribere, F. Solomon, and L. Cote for discussions and reading the manuscript. pAC4 and PBNeoTetO-Dest, the OVCAR4 cells, and THZ531 were gifts from A. Cheng, S. Correa Echavarría, and N. Gray respectively. We thank H. Suzuki for the first stable nucleosome coordinates and F. Lam for assistance with Comet assays. The results shown here are in part based upon data generated by the TCGA Research Network: <http://cancergenome.nih.gov/>. We thank the Koch Institute's Robert A. Swanson (1969) Biotechnology Center at MIT for technical support, specifically G. Paradis of the Flow Cytometry Core and S. Levine of the MIT BioMicro Center. The research described here was supported by Program Project Grant P01-CA042063 from the NCI (P.A.S.), by United States Public Health Service grants R01-GM034277 and R01-CA133404 from the NIH (P.A.S.), and by the Koch Institute Support (core) grant P30-CA14051 from the NCI. S.J.D. was also supported by a David H. Koch Fellowship and NIH Pre-Doctoral Training Grant T32-GM007287 (MIT Biology Department).

Reviewer information Nature thanks R. Fisher, B. Tian and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.J.D., P.L.B. and P.A.S. conceived and designed the experiments and analysis. S.J.D. performed experiments. P.L.B. performed computational analysis. S.J.D., P.L.B. and P.A.S. analysed the data and wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0758-y>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0758-y>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to P.A.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to outcome assessment except in the case of the comet assay results shown in Extended Data Fig. 1g, in which the researcher was blinded to the experimental condition of each sample during image acquisition.

Cell culture, cell line generation and drug conditions. All cell lines were tested for mycoplasma contamination periodically, including immediately upon receipt and after generation of CRISPR-modified clonal cell lines via the MycoAlert Mycoplasma Testing Kit (Lonza). Results were always negative for mycoplasma contamination.

V6.5 (C57Bl/6-129) mES cells and derived cell lines were cultured on 0.2% gelatin-coated tissue culture plates in ES medium: Dulbecco's modified essential medium (Thermo Fisher) buffered with 10 mM HEPES (Thermo Fisher) and supplemented with 15% fetal bovine serum (Hyclone), 1,000 U/ml leukaemia inhibitory factor (Millipore), $1 \times$ non-essential amino acids (Thermo Fisher), 2 mM L-glutamine (Thermo Fisher), 0.11 mM β -mercaptoethanol (Sigma), 100 IU penicillin and 100 μ g/ml streptomycin (Corning). *Cdk12* Δ clones were maintained in 1 μ g/ml doxycycline (Dox) (Sigma) in ES medium (changed daily) to sustain complementing levels of CDK12. To investigate CDK12 loss, cells were washed at time zero with HBS and switched to ES medium without Dox.

Cdk12 Δ clones were generated as follows (see Extended Data Fig. 1a). A *Cdk12*-Flox clone was isolated using CRISPR-Cas9 genome editing technology³¹. Two sgRNA sequences targeting introns 3 and 4 of the endogenous *Cdk12* locus were cloned into pX330³² (a gift from F. Zhang, Addgene# 42230). Lipofectamine2000 (Thermo Fisher) was used to co-transfect wild-type V6.5 cells with the sgRNA plasmids, along with single-stranded oligodeoxynucleotides (ssODNs) from Integrated DNA Technologies (IDT) containing a *LoxP* sequence adjacent to an *NcoI* restriction site flanked on either side by 60-nucleotide homology arms complementary to intron 3 or intron 4 surrounding the sgRNA cut site, and the pLKO.1 plasmid harbouring a puromycin-resistance gene. sgRNA sequences and ssODN sequences are provided in Supplementary Table 5. Cells were selected 24 h after transfection with 1 μ g/ml puromycin (Sigma) for 48 h and single-cell cloned. Clones were screened for homozygous *LoxP* site insertion into both introns by PCR followed by *NcoI* digest. Positive clones were confirmed by Sanger sequencing. A Dox-inducible *Cdk12* transgene was stably introduced into the *Cdk12*-Flox cell line using a piggybac retrotransposon system. N-terminal Flag-HA-tandem epitope-tagged *Cdk12* (NM_001109626.1) was cloned from polyA-selected mouse cDNA into pCR8/GW/TOPO (Thermo Fisher) followed by transfer into the doxycycline-inducible piggybac expression vector, PBNeoTetO-Dest (a gift from A.W. Cheng), using standard TOPO and Gateway cloning kits (Thermo Fisher). This expression vector was cotransfected with pAC4 (constitutively expressing M2rtTA, the Dox-inducible transactivator, flanked by piggybac recombination sites, A.W. Cheng) and mPBase (piggybac transposase expression plasmid, A.W. Cheng) using Lipofectamine2000. 24 h after transfection, cells were selected with 150 μ g/ml Hygromycin (Thermo Fisher) and 200 μ g/ml G418 (Sigma) to select stable transformants. Subsequently, a constitutive Cre expression plasmid, pPGK-Cre-bpA (a gift from K. Rajewsky, Addgene plasmid # 11543), and a constitutive mCherry expression vector, pCAGGS-mCherry³³ were co-transfected using Lipofectamine2000 into the *Cdk12* Flox cells with stably integrated doxycycline-inducible CDK12. 48 h after transfection, the cell population was single-cell FACS sorted for mCherry-positive cells. Beginning 4–6 h after Cre transfection, the cells were treated daily with 1 μ g/ml Dox (Sigma)-supplemented ES medium to express rescuing levels of CDK12 protein. PCR was used to select clones harbouring homozygous deletions of exon 4 and exon 4 deletion was confirmed by Sanger sequencing across the locus. Several homozygous knockout clones were isolated and two clones were picked for subsequent analysis.

Endogenous N-terminal, V5-epitope tagged *Atm*, *Brca2*, and *Fancd2* cell lines were made as follows. sgRNAs targeting genomic loci near the start codon of *Atm*, *Brca2*, and *Fancd2* were cloned into pX330. gBlocks (IDT) containing a V5 epitope tag positioned in-frame, immediately adjacent to the start codon flanked by 64 to 354 nucleotides of homology were TOPO cloned and sequenced verified. sgRNA sequences and gblock sequences are provided in Supplementary Table 5. For *Atm*, where homologous insertion of the V5 tag disrupted sgRNA/Cas9 cutting, we added a restriction enzyme site in frame to the end of the V5 tag to facilitate screening. For *Fancd2* and *Brca2*, where homologous insertion of the V5 epitope tag did not inhibit sgRNA/Cas9 re-cleavage, we also engineered point mutations into the gBlock construct that would introduce a novel restriction enzyme site adjacent to the sgRNA PAM motif to disrupt sgRNA/Cas9 re-cutting and facilitate screening. *Cdk12* Δ cells were co-transfected with the appropriate sgRNA, TOPO-cloned gBlock, and pLKO.1 with blasticidin resistance. Cells were selected with 2 μ g/ml blasticidin and single-cell cloned. Cells were screened by PCR followed by restriction enzyme digest. Heterozygous and homozygous insertions of the V5 tag were isolated and confirmed by Sanger sequencing across the locus. Two independent

clones with homozygous or heterozygous insertions of the V5 tag were isolated and experiments were replicated in at least two independent clones.

22RV1 and PC-3 cells were from ATCC. OVCAR4 cells were from the Koch Institute's High Throughput Sciences Facility Cell Line Repository. All cell lines were authenticated by STR Profiling (ATCC) upon receipt. 22RV1 and OVCAR4 cells were grown in RPMI-1640 (Gibco) supplemented with 10% FBS (Tissue Culture Biologicals), 2 mM L-glutamine (Thermo), and 100 IU penicillin and 100 μ g/ml streptomycin (Corning). PC-3 cells were grown in Ham's F-12K (Kaighn's) medium (Thermo Fisher) supplemented with 10% FBS (Tissue Culture Biologicals), 2 mM L-glutamine (Thermo), and 100 IU penicillin and 100 μ g/ml streptomycin (Corning). Cells were treated with indicated concentrations of THZ531⁴ or equivalent volumes of DMSO (vehicle control) for 4 h before harvest.

FACS assays. FACS analyses were performed using BD FACS machines: FACS Celesta, LSRII, FACS Canto II, FACS LSR Fortessa, and FACS Aria IIIu. Data were collected using FACS Diva Version 8.0.1 and data was analysed using FlowJo version 1.0.1.

Growth curve analysis. 24 h before starting the time course, cells were plated at the same cell number in biological triplicate for each of the first three time points of the experiment (0, 24, and 48 h) in +Dox medium; additional cells were grown in parallel cultures for the later time points. After 24 h, we split these parallel cultures into biological triplicates per condition for the final three time points of the experiment (48, 72, and 96 h). Starting at time 0, cells received daily media changes with +Dox or -Dox medium as appropriate. At each time point, triplicate cell cultures were washed with HBS and harvested by trypsinization. Each biological replicate was resuspended in 450 μ l of ES medium followed by addition of 1 μ l of 50 μ M calcein-AM in DMSO and 2 μ l of 2 mM ethidium homodimer-1 (Thermo Fisher). Samples were incubated for 15–20 min at room temperature protected from light. After staining, 50 μ l of CountBright Absolute Counting Beads (Thermo Fisher) was added to each sample. Samples were analysed by flow cytometry such that at least ~5,000 CountBright Absolute Counting Beads (~100 μ l) were recorded per sample, during which samples were vortexed every minute to prevent counting beads from settling out of solution. The number of live cells per replicate was quantified in each sample by counting the number of live (Calcein-AM positive and Ethidium Homodimer-1 negative) cells and comparing it to the number of CountBright Absolute Counting Beads (with known concentration). An example of the flow cytometry gating strategy used is shown in Supplementary Fig. 1. To calculate the fold change in live cells over the previous 24 h, the number of live cells in each biological replicate at each time point was compared to the average of the live cells in the three biological replicates at the previous time point to give the ratio of live cells every 24 h.

Cell cycle profiling. Cells were plated at approximately equal cell densities 24 h before profiling, such that cells were 50–80% confluent at the time of harvest. Cells were pulsed with 10 μ M 5-ethynyl-2'-deoxyuridine (EdU) for 1 h under standard growth conditions, then harvested by trypsinization. Collected cell pellets were fixed, permeabilized, and stained for EdU incorporation with Alexa Fluor 647 using Click-iT EdU Flow Cytometry Assay Kit (Thermo Fisher) according to the manufacturer's instructions. After EdU staining, cells were resuspended in $1 \times$ Click-iT saponin-based permeabilization and wash reagent (Thermo Fisher) with 50 μ g/ml propidium iodide to label total DNA content and 100 μ g/ml RNase A. Cells were incubated at room temperature in the dark for 30 min, and at least 50,000 cells (Fig. 1c) and 20,000 cells (Extended Data Fig. 1e) gated on P3 (see gating strategy in Supplementary Fig. 1) were analysed by flow cytometry for EdU content (AlexaFluor 647) and total DNA content (propidium iodide). An example of the flow cytometry gating strategy used is shown in Supplementary Fig. 1.

Apoptosis. Cells were plated at approximately equal cell densities at least 24 h before harvest. Cells were harvested by trypsinization. The growth medium and HBS wash before trypsinization were collected and centrifuged with the trypsinized cell population to collect any apoptosing cells with decreased adherence to the plate. Cell pellets were washed twice with cold PBS. Cells were fixed, permeabilized, and stained for cleaved caspase-3 (apoptosis) using the FITC Active Caspase-3 Apoptosis Kit (BD Pharmingen) and the recommended protocol. At least 50,000 stained cells (gated on P3, see gating strategy in Supplementary Fig. 1) were analysed by FACS. An example of the flow cytometry gating strategy used is shown in Supplementary Fig. 1.

Neutral comet assay. Assessment of DNA damage and double strand break formation in cells was performed using the single cell gel electrophoresis assay, CometAssay (Trevigen). Cells were harvested and resuspended in low-melting point agarose, plated onto provided glass slides, lysed, and subjected to electrophoresis in neutral electrophoresis buffer (100 mM Tris, 300 mM sodium acetate, pH 9.0). Slides were processed according to manufacturer's instructions and stained with SYBR Gold. After staining, coverslips were mounted onto slides with approximately 1 drop of ProLong Gold (Thermo Fisher) and cured overnight at room temperature protected from light. DNA tails were visualized using a Nikon

Eclipse 80i fluorescence microscope and quantified using ImageJ software with the OpenComet plugin (<http://www.opencomet.org>)³⁴.

Western blotting. Whole-cell extract was harvested from cells by washing the cells in cold phosphate-buffered saline (PBS) and lysing in RIPA (10 mM Tris pH 7.4, 150 mM NaCl, 1% TritonX-100, 0.1% SDS, 0.5% sodium deoxycholate, and 1 mM EDTA) supplemented with 1 × cComplete, EDTA-free Protease Inhibitors (Roche), 2 μl/ml benzamide nuclease (Sigma), and if needed, 1 × Halt Phosphatase Inhibitor Cocktail (Thermo Fisher). Lysates were incubated on ice for at least 30 min, centrifuged for 10 min at 4 °C and max speed, and the cleared lysate was quantified using a standard BCA assay (Thermo Fisher). Lysates were normalized for equivalent total protein in 1 × loading dye (62.5 mM Tris pH 6.8, 5% glycerol, 2% SDS, 16.67% BME, and 0.083% bromophenol blue) or 1 × NuPAGE LDS Sample Buffer (Thermo Fisher) with 1 × NuPAGE Reducing Agent (Thermo Fisher). Normalized lysates were boiled for 5 min or incubated at 70 °C for 10 min and run on one of the following types of precast gels: NuPAGE 4–12% Bis-Tris gels (Thermo Fisher), NuPAGE 3–8% Tris-Acetate Protein gels (Thermo Fisher), Novex 4–20% Tris Glycine gels (Thermo Fisher), Novex 10–20% Tris Glycine gels. Gels were transferred overnight (30 V) to PVDF in 10% methanol supplemented 1 × NuPAGE Transfer Buffer (NuPAGE Bis-Tris Gels) for Bis-Tris and Tris-Acetate gels or 20% methanol-supplemented 1 × Novex Tris-Glycine Transfer Buffer (Thermo Fisher) for Tris-Glycine gels. Primary antibodies used for blotting: anti-HA high affinity antibody (Roche 11867423001), Enolase I (CST 38105), Vinculin (Sigma V9131), Hsp90 (BD 610418), p53 (1C12) (CST2524S), P-p53 Serine15 (CST 9284S), ATR (CST 13934S), FANCD2 (Abcam ab108928), and V5 (Life Technologies R96025). Secondary antibodies used all blots except the V5 epitope tag: ECL anti-rat IgG (GE Healthcare NA935V), ECL anti-mouse IgG (GE Healthcare NA931V), and ECL anti-rabbit IgG (GE Healthcare NA934V). For blots with the V5 epitope tag, we used the anti-mouse IgG, HRP-linked antibody (CST 7076S). Blots were exposed with Western Lightning Plus-ECL (Perkin Elmer) or SuperSignal West Dura Extended Duration Substrate (Thermo Fisher).

RNA sequencing. Total RNA was harvested using Trizol Reagent (Thermo Fisher) from two independent *Cdk12Δ* clones each in biological duplicate from cells maintained in Dox (+Dox) or withdrawn from Dox for 24 h (–Dox 24 h) or 48 h (–Dox 48 h). In parallel, total RNA was harvested from *Cdk12* floxed cells (without integrated Dox-inducible transgene) that had been pre-treated with 1 μg/ml Dox daily for 17 days and subjected to the same Dox conditions (+Dox, –Dox 24 h, –Dox 48 h) in biological duplicate to serve as a control for gene expression effects of long-term Dox treatment followed by short-term withdrawal. RNA was extracted following the standard Trizol protocol and subsequently DNase treated with Turbo DNase (Thermo Fisher) under standard reaction conditions. RNA quality was assessed by Agilent 2100 Bioanalyzer and only samples with a RIN value ≥ 9 were used for library preparation and sequencing. PolyA-selected libraries were made from 1 μg total RNA input using the TruSeq Stranded mRNA Library Prep Kit (Illumina RS-122-2102) with multiplexing barcodes, following the standard protocol with the following specifications: (1) 5 min RNA fragmentation time, (2) Superscript III (Thermo Fisher) was used for reverse transcription, (3) 15 cycles of PCR were used during the library amplification step, and (4) AMPure beads (Beckman Coulter) were used to size select/purify the library after PCR amplification instead of gel size selection. The 18 libraries were pooled and sequenced on two lanes of an Illumina NextSeq500.

RT-qPCR. For total RNA (THZ531 treatments of human cell lines): cells were harvested in Trizol (Thermo Fisher) following the manufacturer's instructions. Isolated RNA was DNase treated with Turbo DNase (Thermo Fisher) under standard reaction conditions. Subsequently, RNA was extracted with acid phenol:chloroform and ethanol precipitated. Reverse transcription was performed using SuperscriptIII (Thermo Fisher) with normalized total RNA input (4–5 μg per reaction) in a 20-μl reaction with 1 μl of 50 μM olido(dT)₂₀ primer and standard reaction conditions (50 °C, 1 h). Each cDNA reaction was treated with 5U RNase H at 37 °C for 20 min and subsequently diluted 1:10 in ddH₂O. Real-time qRT-PCR was performed using Power SYBR Green (Thermo Fisher) with 1–4 μl 1:10 diluted cDNA (depending on the target) and 200 nM forward and reverse primer mix (500 nM for *HPRT* control primer pair) on an ABI 7500 Real-Time qPCR machine or an ABI StepOnePlus machine. Primers spanning the exon 3–exon 4 junction of *HPRT* were used as a control to compare samples using a standard $\Delta\Delta Ct$ calculation. See Supplementary Table 6 for primer sequences.

For purified mRNA (qPCR from *Cdk12Δ* clones): total RNA was purified directly from cells using the RNeasy Plus kit (Qiagen) and following the manufacturer's instructions. Subsequently, mRNA was purified from total RNA using the Oligotex mRNA Mini kit (Qiagen). Reverse transcription was performed using SuperscriptIII (Thermo Fisher) with normalized quantities of mRNA input (200–250 ng per reaction) in a 20-μl reaction with 200 ng random hexamer primers and standard reaction conditions (50 °C, 1 h). Each cDNA reaction was treated with 5U RNase H at 37 °C for 20 min and diluted 1:10 in ddH₂O. Real-time qRT-PCR was performed using Power SYBR Green (Thermo Fisher) with 1–4 μl 1:10 diluted

cDNA (depending on target) and 200 nM forward and reverse primers on an ABI 7500 Real-Time qPCR machine or an ABI StepOnePlus machine. Primers to *Pgk1* were used as a control to compare samples using a $\Delta\Delta Ct$ calculation. See Supplementary Table 6 for primer sequences.

Chromatin immunoprecipitation sequencing. We modified a high-resolution, micrococcal nuclease (MNase) digestion-based ChIP methodology^{35,36}. We performed ChIP on *Cdk12Δ* cells under +Dox or –Dox 48 h conditions for total RNAPII density and Ser2p RNAPII density. Two independent antibodies were used for each antigen as follows: 8WG16 (Abcam ab817) and Rpb3 (Bethyl A303-771A) for total RNAPII density; and H5 Clone (Abcam ab24758) and 3E10 Clone (Millipore 04-1571) for Ser2p RNAPII density. Two biological replicates were processed for each antibody and Dox condition (for example of the experimental setup for one protein target, for example, RNAPII, see Extended Data Fig. 4). For each Dox condition, we also processed four negative control libraries: one whole-cell extract (WCE) sample and three mock immunoprecipitation samples with the following antibodies: goat anti-mouse IgM (Thermo Fisher 31172), goat anti-rat IgG (Thermo Fisher 31226), mouse IgG2a [MOPC-173] isotype control (Abcam ab18413).

Chromatin immunoprecipitation was performed as follows. In brief, 48 h before cells were harvested, 10⁷ *Cdk12Δ* cells were plated in 15 cm dishes in either +Dox or –Dox ES medium. Cells were crosslinked directly in medium on the plate in 1% methanol-free formaldehyde (Thermo Fisher) for 10 min at room temperature. Crosslinking was quenched with 250 mM glycine. Cells were washed 3 × in 10 ml chilled PBS and harvested by scraping. Cells were pelleted, washed in 10 ml chilled PBS, and pelleted again. PBS was aspirated off the cells and the pellets were flash frozen in liquid nitrogen and stored at –80 °C. Pellets were thawed on ice and resuspended in 0.9 ml of ChIP lysis buffer (1% SDS, 10 mM EDTA, 50 mM Tris-HCl pH 7.5) supplemented with 1 × cComplete Protease Inhibitors (Roche) and 1 × Halt Phosphatase Inhibitors (Thermo Fisher) and lysed for 10 min at room temperature. Pellets were diluted in 8.1 ml of ChIP Dilution Buffer (1% TritonX-100, 2 mM EDTA, 150 mM NaCl, 20 mM Tris-HCl pH 7.5) supplemented with 1 × cComplete Protease Inhibitors (Roche) and 1 × Halt Phosphatase Inhibitors (Thermo Fisher) and 3 mM CaCl₂. Each tube was pre-warmed at 37 °C for 2 min. Twelve units of micrococcal nuclease (MNase, Sigma) were added per tube and samples were digested for 30 min at 37 °C with rotation. The digestion reaction was quenched by the addition of 180 μl 500 mM EDTA and 360 μl 500 mM EGTA per tube. Samples were sonicated in 15-ml polystyrene tubes in a BioRupter (Diagenode) for 20 cycles on high (1 cycle = 30 s on/30 s off). Samples were cleared by centrifugation (max speed, 4 °C, 10 min).

Soluble material was transferred to a new tube and each sample (one pellet) was split into four 1.8 ml parts. Lysate was incubated overnight at 4 °C with rotation with 10 μg or 10 μl of acetic acid of the following antibodies: total RNAPII (8WG16 and Rpb3), and RNAPII CTD Ser2p (H5 and 3E10). After the overnight incubation, the immunoprecipitates were incubated for 2 h at 4 °C as follows: the 8WG16 and Rpb3 immunoprecipitates were incubated with 100 μl of Protein G Dynabeads (Thermo Fisher). The H5 immunoprecipitates were incubated with 100 μl of Protein G Dynabeads pre-conjugated overnight with 20 μg goat anti-mouse IgM (Thermo Fisher 31172) in Bead Preparation Solution (9 ChIP Dilution Buffer: 1 ChIP Lysis Buffer). The 3E10 immunoprecipitate was incubated with 100 μl of Protein G Dynabeads pre-conjugated with 20 μg goat anti-rat IgG (Thermo Fisher 31226). Immunoprecipitates were washed as follows: 2 × 2 ml LB3 (20 mM Tris-HCl pH 7.5, 150 mM NaCl, 2 mM EDTA, 0.1% SDS, 1% Triton X-100); 1 × 2 ml LB3+ (20 mM Tris-HCl pH 7.5, 500 mM NaCl, 2 mM EDTA, 0.1% SDS, 1% Triton X-100); 1 × 2 ml Lithium Chloride Buffer (10 mM Tris-HCl, pH 7.5, 250 mM LiCl, 1 mM EDTA, 1% NP-40), and 1 × 2 ml TE + 50 mM NaCl (10 mM Tris-HCl, pH 7.5, 1 mM EDTA, 50 mM NaCl). After washing, beads were resuspended in 200 μl of extraction buffer (50 mM Tris pH 8, 10 mM EDTA, 5 mM EGTA, 1% SDS). For WCE control, 100 μl of soluble input material was added to 100 μl of extraction buffer. Samples were incubated overnight at 65 °C with 1,000 r.p.m. shaking to elute immunoprecipitates and reverse crosslinks. 200 μl of eluted samples was cleared of beads and added to 200 μl TE pH 8. Samples were digested with 0.2 mg/ml RNaseA for 1 h at 37 °C. 7 μl of CaCl₂ solution (10 mM Tris-HCl, pH 8 and 300 mM CaCl₂) and 4 μl of 20 mg/ml proteinase K were added to each sample and incubated for 1 h at 55 °C. Samples were extracted in 1 × phenol:chloroform followed by 1 × chloroform and precipitated overnight at –80 °C with a standard NaCl, ethanol, and glycogen DNA precipitation. Pellets were washed twice in 1 ml 70% ethanol, dried, and resuspended in 70 μl 0.1 × TE pH 8.

All of the ChIP material (70 μl resuspended) or 200 ng of input material (WCE) was used to prepare libraries. Sample DNA was end-repaired for 30 min at 20 °C in a 100 μl reaction: 1 × T4 DNA Ligase Buffer (New England Biolabs, NEB), 0.4 mM dNTPs (NEB), 15 U of T4 DNA Polymerase (NEB), 5 U of Klenow enzyme (NEB), 50 U of T4 PNK (NEB). Samples were purified by Invitrogen PureLink Kit (Thermo Fisher) following standard conditions and eluted in 33 μl. 32 μl of purified product was A-tailed for 37 °C for 30 min in a 50-μl reaction with 1 × NEB Buffer 2, 2 μM

dATP (NEB), and 15 U of Klenow exo- (NEB). Samples were purified by Invitrogen PureLink Kit and eluted in 11 µl. Illumina genomic adapters were ligated onto 10 µl of the purified product for 15 min at 20 °C in the following 50-µl reaction with 1 × Quick Ligase Buffer (NEB), 400 nM of Y-shaped Adaptor Oligo, and 5 µl of Quick Ligase. The reaction was cleaned up by double size selection with AMPure beads as described by the manufacturer with the initial size selection using a 0.9 × AMPure ratio and keeping the supernatant (to select against large products: mononucleosomes and larger) followed by a 1.8 × AMPure selection keeping the bead bound material (to select against adaptor dimers and free adaptors). Note: when switching between first and second size selection, we used the following formula provided by the manufacturer to calculate the amount of additional beads to add: (second ratio – first ratio) × volume transferred. Illumina adaptor oligos were added to the size selected product in a 50 µl PCR reaction with 200 µM dNTPs (NEB), 1 × High Fidelity Phusion Buffer (NEB), 1 µl Phusion Polymerase (NEB), 0.5 µM forward Illumina oligo adaptor, and 0.5 µM reverse Illumina oligo adaptor with 16 × cycles of standard Phusion PCR conditions (annealing: 65 °C 30 s and extension: 72 °C 30 s). PCR products were AMPure purified (1.8 × ratio) according to manufacturer guidelines. Libraries were run on the BioAnalyzer. Libraries with adaptor dimer contamination were repurified with extra AMPure selections until adaptor dimer contamination disappeared. The libraries were sequenced with 40 bp paired-end reads on 4 lanes of an Illumina HiSeq 2000. Sequencing results from the four lanes were pooled and analysed.

Bioinformatics and statistical analyses. *Mapping reads and junctions.* Raw RNA-seq reads were mapped using Tophat v. 2.0.12³⁷ or STAR aligner version 2.4.1d³⁸ with the parameters:

```
STAR --runMode alignReads --runThreadN 2 --genomeDir UCSC_mm9 --two-passMode Basic --sjdbOverhang 74 --outSAMtype BAM SortedByCoordinate --outFilterMultimapNmax 20 --outFilterMismatchNmax 999 --outFilterMismatchNoverLmax 0.04 --alignIntronMin 70 --alignIntronMax 500000 --alignMatesGapMax 500000 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --outSAMstrandField intronMotif --outFilterType BySJout
```

Gene expression quantification. Gene-level quantification was performed using Rsem version 1.2.26³⁹ and EBSeq version 1.10.1⁴⁰ using the UCSC mm9 genome annotation with four independent replicates each (2 × technical replicates of 2 × independent clones) for CDK12-expressing and CDK12-depleted samples at 24 and 48 h post Dox-withdrawal. Parameters:

```
rsem-calculate-expression --forward-prob 0.5 --output-genome-bam -p 2 --paired-end UCSC_mm9
```

Rsem analysis was followed by read-count matrix assembly and standard EBSeq differential expression analysis.

Gene-level and quantification was also performed using DESeq⁴¹ to obtain independent validation as well as to derive an adjusted *p* value for use in gene expression volcano plots. Gene expression differences at the total gene level were considered significant if the PPDE as determined by EBSeq > 0.95. Genes that exhibited statistically significant changes in response to doxycycline alone were excluded from the set of CDK12-affected genes. Gene expression data are provided in Supplementary Table 1.

Annotation of p53-target genes and bivalent promoters. p53 target genes were derived from previously published data²⁰. In order to be considered a direct p53 transcriptional target gene, the gene was required to have a p53 ChIP peak within ~5.5 kb upstream and 2.5 kb downstream of the TSS, and exhibit a p53-dependent transcriptional response to DNA damage induced by adriamycin (doxorubicin) dosing in mouse ES cells. The directionality of transcriptional responses was determined by microarray-assayed gene expression changes.

Classification of bivalent-promoter genes was based on previously published data²¹. Genome-wide ChIP experiments were used as the basis for classifying genes based on histone modifications within the promoter region. Genes with H3K4-trimethylation overlapping H3K27-trimethylation in mouse ES cells were considered to belong to the bivalent class.

Determination and quantification of alternative splicing events. Mapped splice junctions detected in all samples were combined and processed using custom Python scripts to filter out junctions representing < ~1% of transcripts. Alternative and constitutive intron classifications were performed using custom Python scripts, and are agnostic with regard to existing annotations other than known gene boundaries. If no overlapping introns exist for a given intron, it is assigned to the constitutive class. The subgroups containing overlapping introns are assigned a splicing classification if the start and end coordinates of all of the constituent introns fall into a pattern representing a known splice type (cassette, mutually exclusive, alternative 5' splice site, alternative 3' splice site).

To quantify differences in alternative splicing events between CDK12-expressing and CDK12-depleted cells, splicing events determined from mapped junctions as described above were converted into event-specific gff3 annotation files compatible with MISO⁴². We then performed two separate MISO analyses (one for each clone, comparing CDK12-expressing versus CDK12-depleted samples) per alternative

splicing type. To be considered significant, splicing events in both independent clones were required to change in per cent-spliced in (delta-PSI) in the same direction, both with a Bayes factor of ≥ 5 .

Annotation of IPA sites. We used previously published 3' end sequencing data²⁵ to identify genome-wide polyadenylation sites in the same strain of mouse ES cells used to generate the *Cdk12*Δ lines. These data were generated from poly(A)-selected RNA that was oligo-dT primed and reverse transcribed. cDNAs were circularized, PCR amplified, and sequenced. After mapping, the reads were filtered to remove genomically encoded poly(A) tracts and polyadenylation sites associated with B2-SINE retrotransposons. Putative cleavage sites were then required to have one of 36 strong, polyadenylation signal sequences within 40 nucleotides upstream. In the data analysis performed here, we first combined all cleavage sites from both control and U1-snRNA antisense morpholino oligonucleotide treated cells (two replicates each) and cleavage sites within 40 nucleotides of contiguous sequence were combined into the same cluster. Clusters were required to contain a minimum of 20 reads to be included in the analysis. Putative IPAs that overlapped an alternative 5' splice site extending into the intron (as identified in our RNA sequencing data; see 'Determination and quantification of alternative splicing events' above) were excluded from the analysis. The genomic coordinates of these clusters are provided in Supplementary Table 2.

Genome-wide IPA transcript annotation and quantification. We used custom Python scripts to derive a transcriptome annotation based on Mm9 (Mus_musculus_NCBI_build37.1) gene start and end boundaries, the location of polyadenylation sites as determined by 3' end sequencing (described above), and the genomic locations of all mapped splice junctions from the RNA-seq data. Annotation of the distal polyadenylation site isoform for each gene was based on the consensus isoform, that is, the junctions defining each exon are the most frequently used junction detected among all of the combined samples. Each IPA site within a gene was then assigned to an additional transcript based on the consensus isoform but terminating at the IPA cleavage site. These gene annotations were then converted to DEXseq exon parts using DEXseq-associated script `dexseq_prepare_annotation.py` and the reads mapping to each exon part in each sample were counted using `dexseq_count.py`. Using the counts matrix thus derived for all biological replicates in each condition, DEXseq was used to identify changes in the relative abundance of each exon part as normalized to all exon parts within the gene, including those representing the IPAs and distal polyadenylation site isoform 3' terminal exon. This gave a log₂-fold change and FDR adjusted *P* value for each exon part as it differed between the CDK12-expressing and CDK12-depleted samples. IPA sites or distal polyadenylation site isoform 3' terminal exons whose exon part exhibited an adjusted *P* value < 0.05 were considered statistically significant. Significantly changing IPA sites and distal polyadenylation site isoform 3' terminal exons are listed in Supplementary Tables 3 and 4, respectively.

Determination of first stable nucleosome dyad positions. Genomic coordinates of nucleosome dyads in mES cells were download from previously published data⁴³. To identify regions with stable nucleosomes, five different mES cell MNase-seq data sets were analysed with NucTools⁴⁴. Stable nucleosome regions were determined using `stable_nucs_replicates.pl` with a sliding window of 50 bases and a relative error based on five replicates < 0.5. The NucTools-defined stable region dyad downstream and most proximal to the transcription start site was regarded as the +1 dyad (first stable nucleosome).

ChIP sequencing analysis. Custom Python and R scripts were used to calculate normalized read densities of ChIP data, bin for metagene analysis, and perform statistical tests. In brief, genomic coordinates for full genes or TSS-flanking regions for specific gene sets were first compiled. Mapped reads from ChIP data sets were counted within the specified regions using the `Bedtools coverageBed` tool⁴⁵, combining both replicates of both antibodies for the ChIP under analysis. Each gene was divided into 100 equal-length bins and the total read counts per nucleotide for each gene within each bin were summed. Summed read counts were normalized by total mapped reads for the sample under consideration and the average of these count values across all genes was plotted as the normalized read density for that bin. Bin-wise *P* values were obtained using a one-sided Kolmogorov-Smirnov test comparing the distributions of normalized reads across all genes in the group between CDK12⁺ and CDK12⁻ samples (see Extended Data Fig. 4).

Determination of Cdk12 sensitivity index. In order to determine the cumulative effect of IPA site usage on full-length gene expression, the log₂-fold change between CDK12-depleted and CDK12-expressing samples in the ratios of the custom annotated first and last (distal polyA isoform) exons of each gene was calculated. We refer to this change in ratio as the CDK12 sensitivity index. The index for each gene was plotted as part of a distribution containing all other genes sharing the same total number of detected IPA sites as determined by the IPA annotation (see above).

Identification and quantification of IPA sites in TCGA tumour samples. Patient samples from the prostate adenocarcinoma²⁷ and ovarian serous cystadenocarcinoma²⁹ TCGA cohorts were assessed for the presence of missense or truncating point mutations as well as copy-number variations (amplifications or deletions) in

CDK12 and *BRCAness* genes using cBioPortal (<http://www.cbioportal.org/>)^{46,47}. Additionally, normalized *CDK12* mRNA expression levels were considered. We included all tumours from these cohorts in our analysis that were predicted to have *CDK12* loss-of-function (LOF) mutation(s) as annotated in cBioPortal. We considered tumour genotypes to be likely *CDK12* LOF mutations if they carried at least one truncating or missense putative driver mutation or if the copy number analysis classified the tumour as carrying a deep deletion (indicating a likely homozygous deletion across the locus) and if the mRNA expression levels of *CDK12* were significantly downregulated compared to the mean expression in wild-type, diploid *CDK12* tumours. We additionally included 4–5 shallow deletions from each tumour type that exhibited the lowest mRNA expression levels of *CDK12*, and one in-frame deletion mutant of unknown functional consequence. As an additional control for specificity, we included a single tumour carrying a *CDK12*(K975E) mutation that had been previously validated as a missense mutation with no LOF. The only *CDK12* mutated tumours we excluded from consideration were three tumours in prostate carrying missense mutations of unknown consequence as we could not accurately classify them as wild-type or LOF.

A set of patients from the prostate and ovarian cohorts with wild-type, diploid *CDK12* loci were selected along with one or two samples from each tumour type carrying an amplified locus. Among all such tumours, this control set was selected by ranking the tumours in order of normalized *CDK12* mRNA expression and taking the subset with the highest expression (9 from each cohort). For the *BRCAness* control subset, we selected a set of tumours that carried only 'probably oncogenic' missense or truncating mutations and selected a set that contained all available *BRCAness* genes, as well as larger samples for genes that are more frequently mutated (for example, *BRCA1/2* and *CHEK2*). These tumours were selected without considering *CDK12* gene expression levels. Retrospectively, we also determined that two of the *CDK12* wild-type diploid tumours carried putative deep deletions with mRNA loss in *FANCA*, *CHEK1*, and *CHEK2* in one tumour and *BRCA2* in the other; these tumours exhibited identical low IPA usage consistent with the other *CDK12* wild-type tumours. Once the tumour sets had been selected based on these genomic characteristics, sequencing data from all of these tumours and only these tumours were downloaded and included in the quantifications performed as described below.

Aligned reads (STAR-mapped.bam files) covering relevant gene loci were downloaded from the Genomic Data Commons (<https://gdc.cancer.gov/>) using the bam-slicing tool. Reads were visually inspected to identify regions with clear IPA events that aligned with a canonical PAS motif. Regions spanning from the upstream 5' splice site of the preceding exon to the PAS site were added to gtf annotations of the gene of interest, and DEXseq tools were used for annotating and counting reads mapping to each exon part as described above; these were used to generate a matrix of counts per exon part in each individual tumour sample. All exonic reads aligning within the gene were summed for each sample, and then read counts for each exon part were normalized to exon length and total exon counts in the gene for each sample to obtain a normalized IPA site usage metric for that sample. GraphPad PRISM 7 software was used to generate plots. Significance levels of the differences in IPA site usage between wild-type diploid versus the deletion and mutation group were determined using the Mann–Whitney *U* test (one-tailed). **Statistics and reproducibility.** All results with one representative example shown were repeated at least twice ($n = 2$ independent experiments) with each of those independent experiments from two independently derived clonal cell lines. Western blots for *CDK12* expression: *Cdk12Δ* clone 36 (Fig. 1a) $n = 4$ independent experiments and *Cdk12Δ* clone 28 (Extended Data Fig. 1c) $n = 4$ independent experiments. Cell cycle profiling: *Cdk12Δ* clone 36 (Fig. 1c) $n = 3$ independent experiments and *Cdk12Δ* clone 28 (Extended Data Fig. 1e) $n = 2$ independent experiments. Cleaved Caspase 3 Apoptosis Staining: *Cdk12Δ* clone 36 (Fig. 1d) $n = 2$ independent experiments and *Cdk12Δ* clone 28 (Extended Data Fig. 1f) $n = 2$ independent experiments. Comet Assays: *Cdk12Δ* clone 36 (Fig. 1e) $n = 1$ independent experiment and *Cdk12Δ* clone 28 (Extended Data Fig. 1g) $n = 1$ independent experiment. Western blots for p53 and Ser15 phosphorylated p53: *Cdk12Δ* clone 36 (Fig. 1f) $n = 2$ independent experiments and *Cdk12Δ* clone 28 (Extended Data Fig. 1h) $n = 1$ independent experiment. *Cdk12Δ* genotyping gel (Extended Data Fig. 1b) $n = 2$ independent experiments and

PCR products were sequenced verified by Sanger sequencing. Endogenous ATR and FANCD2 expression upon *CDK12* loss (Fig. 4d) were repeated in $n = 2$ independent experiments. All V5-epitope tagged western blots (Fig. 4e) were repeated in two independently derived homozygously or heterozygously V5-epitope, endogenously tagged clones. V5-ATR western blot (Fig. 4e): clone 1 ($n = 1$) and clone 7 ($n = 2$). V5-BRCA2 (Fig. 4e): clone 14 ($n = 2$) and clone 15 ($n = 2$). FANCD2 (Fig. 4e): clone 5 ($n = 3$) and clone 35 ($n = 3$). Western blots from human cell lines treated with THZ531 (Extended Data Fig. 10c) were performed in $n = 1$ independent experiments to establish the concentration at which 22RV1, PC-3, and OVCAR4 cells responded to THZ531 treatment. qRT-PCR experiments based on these treatments (Extended Data Fig. 10d, e) were performed in $n = 3$ biological replicates.

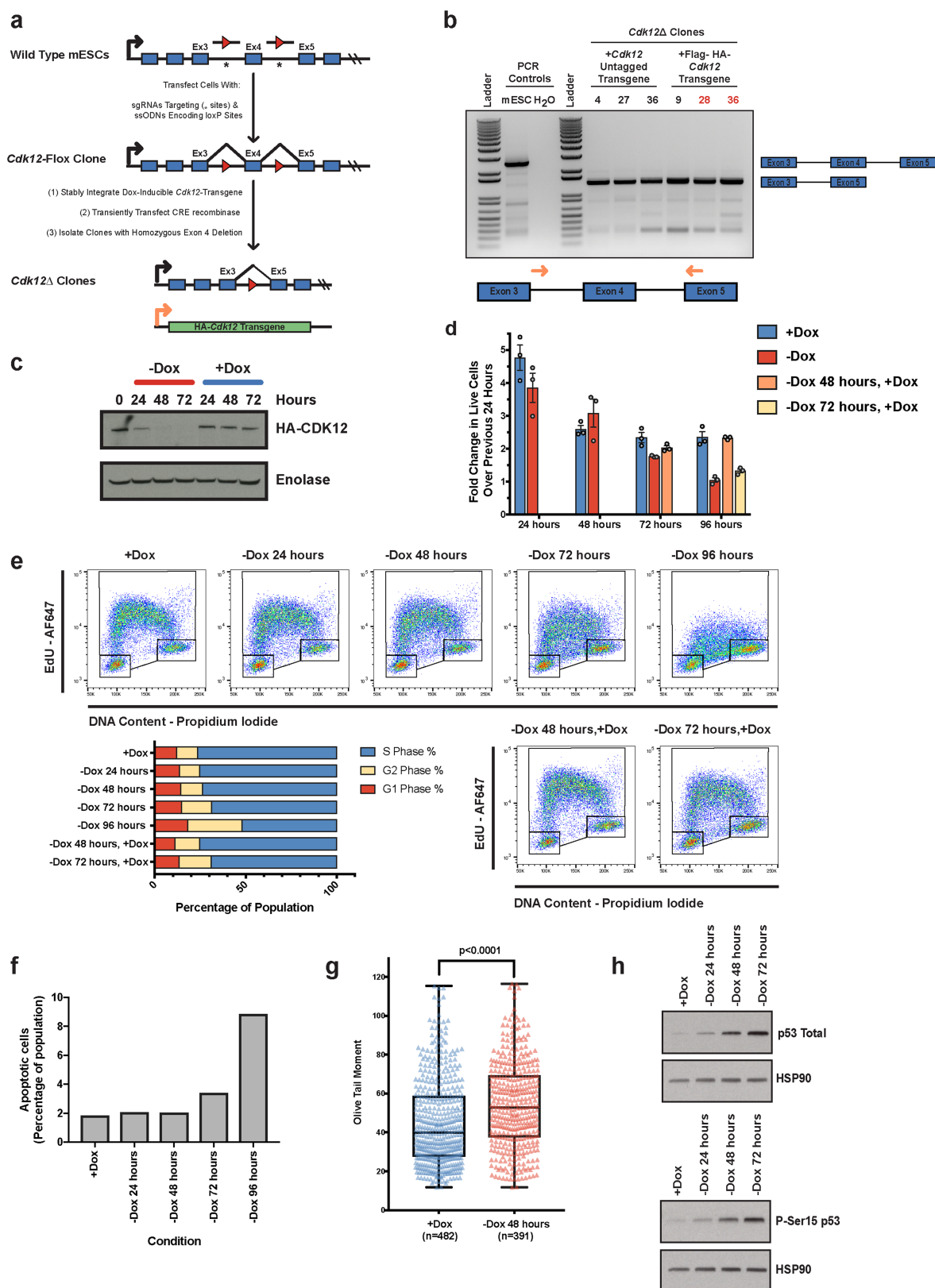
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. Custom scripts used for analysis are available upon request.

Data availability

Sequencing data have been deposited in the Gene Expression Omnibus under accession number GSE116017.

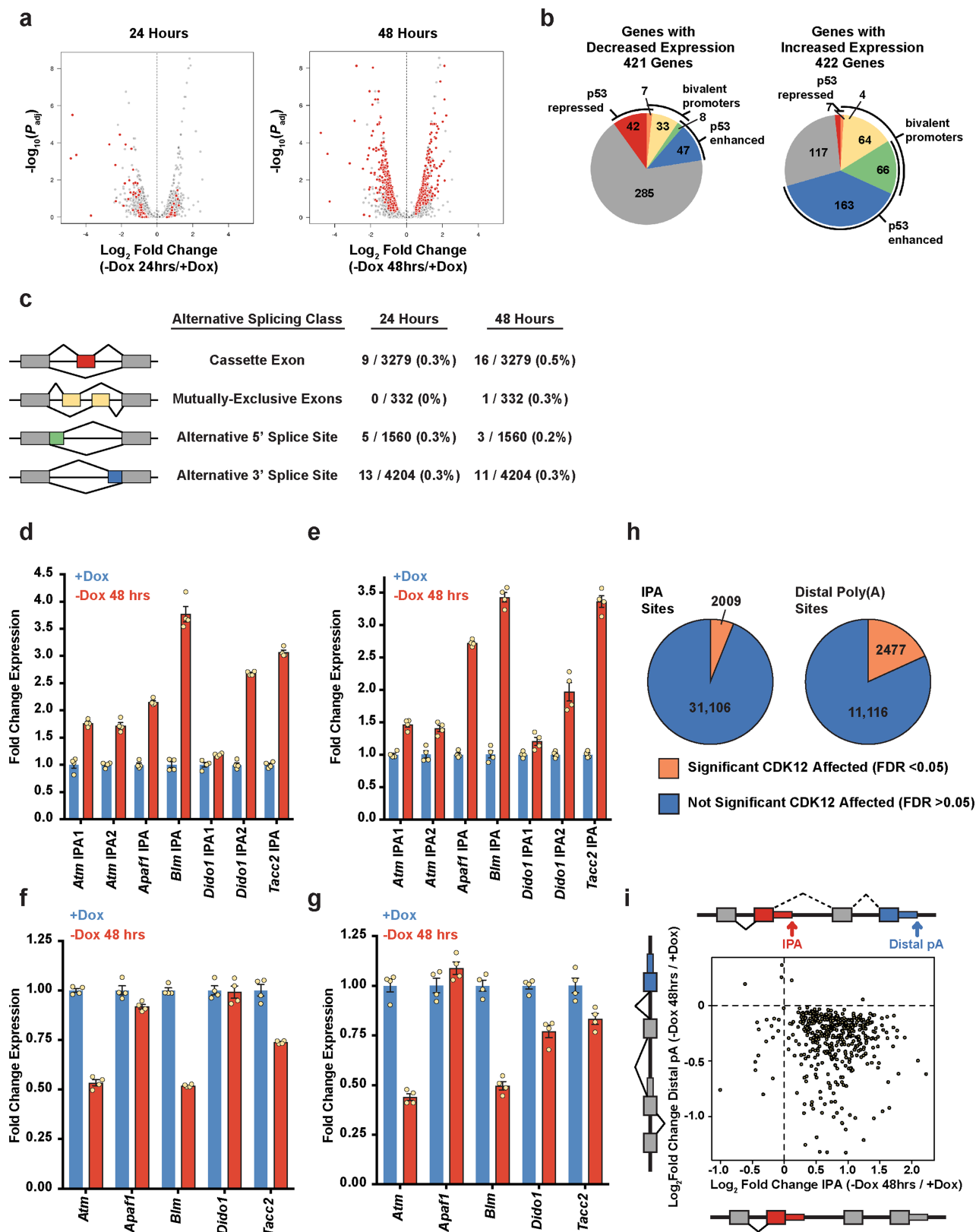
- Ran, F. A. et al. Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
- Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
- Gurtan, A. M., Lu, V., Bhutkar, A. & Sharp, P. A. In vivo structure-function analysis of human Dicer reveals directional processing of precursor miRNAs. *RNA* **18**, 1116–1122 (2012).
- Gyori, B. M., Venkatachalam, G., Thiagarajan, P. S., Hsu, D. & Clement, M.-V. OpenComet: an automated tool for comet assay image analysis. *Redox Biol.* **2**, 457–465 (2014).
- Skene, P. J. & Henikoff, S. A simple method for generating high-resolution maps of genome-wide protein binding. *eLife* **4**, e09225 (2015).
- Skene, P. J., Hernandez, A. E., Groudine, M. & Henikoff, S. The nucleosomal barrier to promoter escape by RNA polymerase II is overcome by the chromatin remodeler Chd1. *eLife* **3**, e02042 (2014).
- Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
- Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
- Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
- Leng, N. et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* **29**, 1035–1043 (2013).
- Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
- Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–1015 (2010).
- Voong, L. N. et al. Insights into nucleosome organization in mouse embryonic stem cells through chemical mapping. *Cell* **167**, 1555–1570.e15 (2016).
- Vainshtein, Y., Rippe, K. & Teif, V. B. NucTools: analysis of chromatin feature occupancy profiles from high-throughput sequencing data. *BMC Genomics* **18**, 158 (2017).
- Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
- Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013).
- Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
- Cui, Y. & Denis, C. L. In vivo evidence that defects in the transcriptional elongation factors RPB2, TFIIS, and SPT5 enhance upstream poly(A) site utilization. *Mol. Cell. Biol.* **23**, 7887–7901 (2003).
- Yang, Y. et al. PAF complex plays novel subunit-specific roles in alternative cleavage and polyadenylation. *PLoS Genet.* **12**, e1005794 (2016).
- Liu, X. et al. Transcription elongation rate has a tissue-specific impact on alternative cleavage and polyadenylation in *Drosophila melanogaster*. *RNA* **23**, 1807–1816 (2017).
- Bentley, D. L. Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.* **15**, 163–175 (2014).



Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Generation of *Cdk12* genetic knockouts in mES cells and phenotypic data from a second, independently derived *Cdk12Δ* clone. **a**, Schematic of *Cdk12Δ* cell line generation, LoxP sites (red triangles), sgRNA cut sites (*), endogenous promoter (black arrows) and doxycycline-inducible promoter (orange arrow). **b**, PCR products across the *Cdk12* locus flanking exon 4 (primers shown as orange arrows) for wild type mES cells and *Cdk12Δ* clones. Clones 28 and 36 used throughout this study are indicated in red. **c–h**, Phenotypic data from the second of two independently derived *Cdk12Δ* clones shown corresponding to results shown in Fig. 1a–f for the other clone. **c**, Representative immunoblot for CDK12 transgene (HA epitope) expression after doxycycline (Dox) withdrawal. **d**, Fold-change in live cells over previous 24 h quantified by FACS; bars represent mean fold change (\pm s.e.m.) for $n = 3$ biological replicates. Cells were grown continuously in

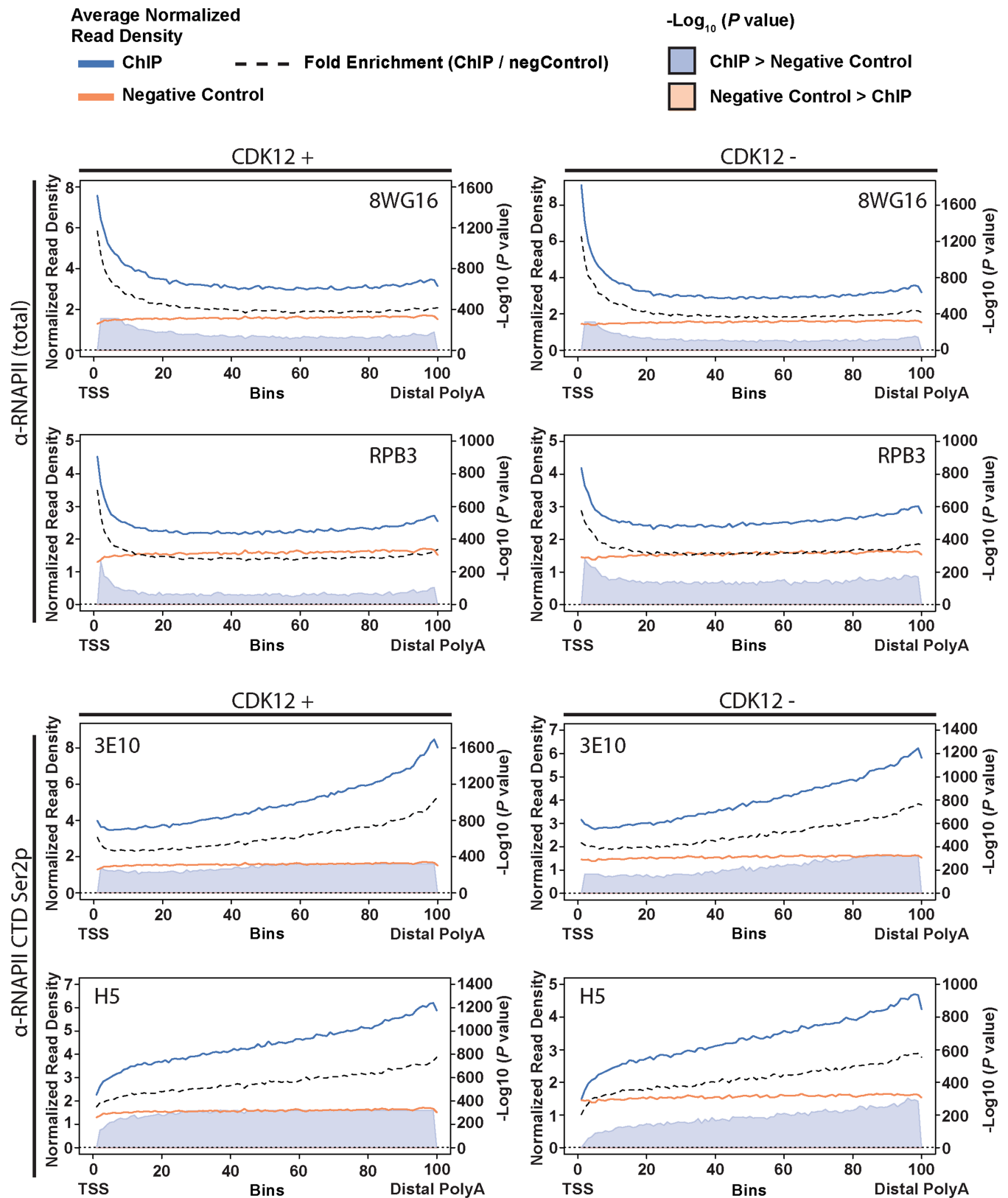
Dox (blue bars), withdrawn from Dox at time 0 and maintained off Dox (red bars), or withdrawn from Dox at time 0 and reintroduced to Dox after 48 h (orange bars) or 72 h (yellow bars) for remainder of the time course. **e**, FACS-based cell cycle profiling of one representative replicate for the same conditions as in **d** (top) and quantification (bottom). **f**, FACS-based quantification of cleaved caspase 3-positive (apoptotic) cells. One representative biological replicate shown. **g**, Neutral comet assay to quantify degree of unrepaired DNA double-stranded breaks in *Cdk12Δ* cells after 48 h of doxycycline withdrawal. Boxplots: median value with 25th and 75th quartiles; whiskers show minimum to maximum. *P* value based on one-sided Mann–Whitney *U* test. **h**, Immunoblot for total and phosphorylated Ser15 (P-Ser15) p53 upon CDK12 loss for the indicated times. HSP90 serves as a loading control.



Extended Data Fig. 2 | See next page for caption.

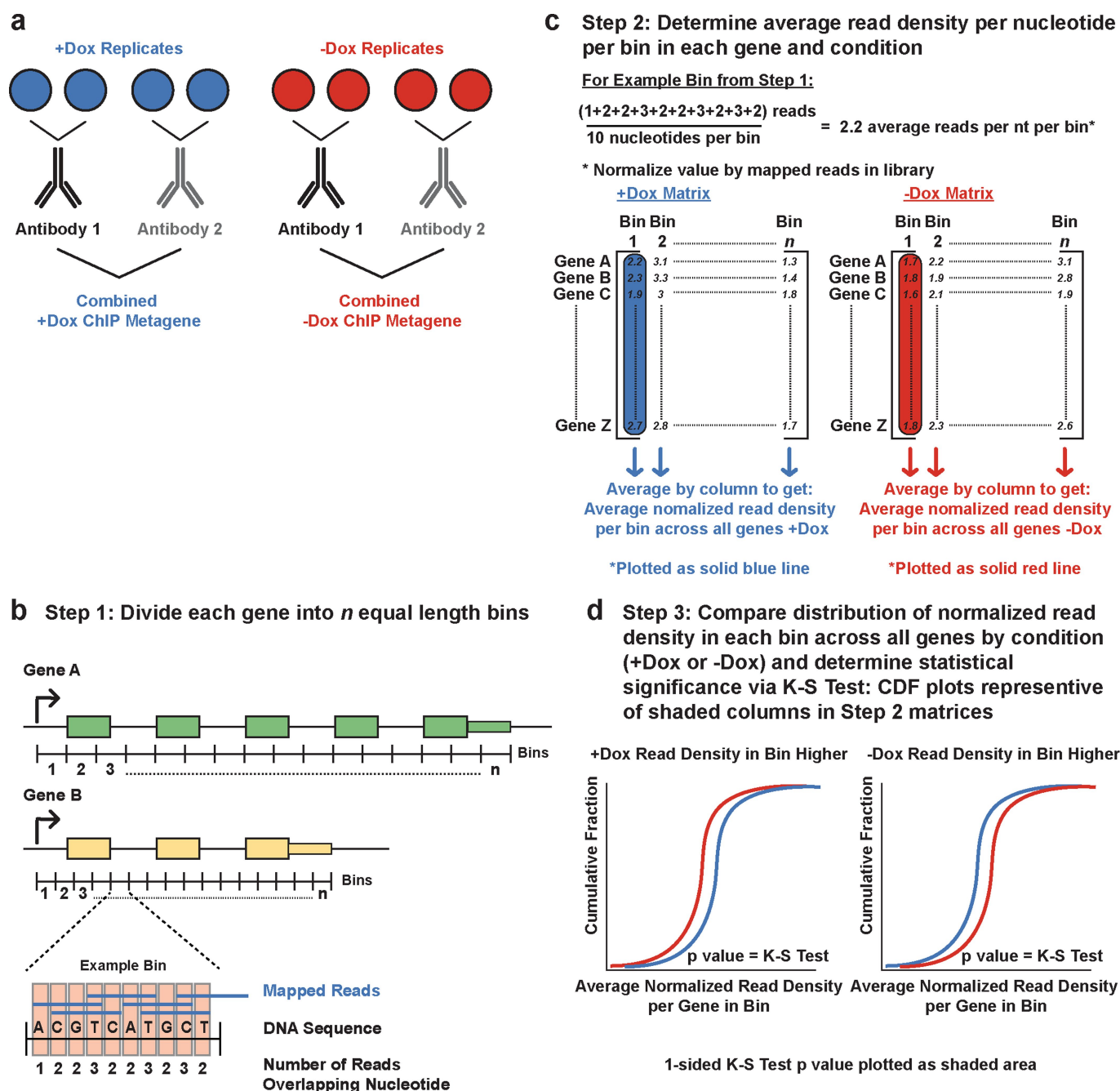
Extended Data Fig. 2 | Gene expression changes in CDK12-depleted mES cells are dominated by increased IPA usage. **a**, Volcano plots of significant gene expression events at the total gene level after 24 h (left) or 48 h (right) of Dox withdrawal. *y*-axis: FDR-adjusted *P* value determined by the DESeq package in R; coloured dots: PPDE > 0.95 (determined by the EBSseq package in R). **b**, Pie chart indicating genes that decreased (left) or increased (right) in total gene expression at a statistically significant level after 24 or 48 h of Dox withdrawal (combined). Likely secondary effects are indicated: p53 repressed genes (red), p53 enhanced genes (blue), bivalent promoter genes (yellow), p53 repressed and bivalent promoter genes (orange), and p53 enhanced and bivalent promoter genes (green). Genes belonging to none of the above groups are indicated in grey. **c**, Table summarizing significant alternative splicing events observed after 24 and 48 h of Dox depletion in *Cdk12*Δ cells. **d–g**, Isoform-specific RT–qPCR corroborating differential isoform usage observed in the RNA

sequencing data. Blue bars (+Dox) and red bars (–Dox 48 h) represent mean (± s.e.m.), *n* = 4 biological replicates. Seven IPA isoforms from five genes were validated in two independent *Cdk12*Δ clones in **d**, **e** and the corresponding distal polyadenylation isoforms in those five genes were validated in **f**, **g**. **d**, **f** and **e**, **g** represent corresponding data from the two independently derived *Cdk12*Δ clones used throughout this study. **h**, Left, IPA sites exhibiting a statistically significant ($P_{\text{adj}} < 0.05$, FDR adjusted *P* value determined by the DEXSeq package in R) change (orange) or not (blue) in expressed genes after 24 or 48 h of Dox depletion. Right, expressed genes with terminal polyadenylation sites that are significantly changed (orange) or not statistically significant (blue) as normalized to the rest of the transcript. **i**, Scatterplot showing log₂ fold-change upon CDK12 loss in distal exons (*y*-axis) versus IPA sites (*x*-axis) in genes that have both a statistically significant ($P_{\text{adj}} < 0.05$) IPA site and a statistically significant distal polyadenylation change; *n* = 4 biological replicates per condition.



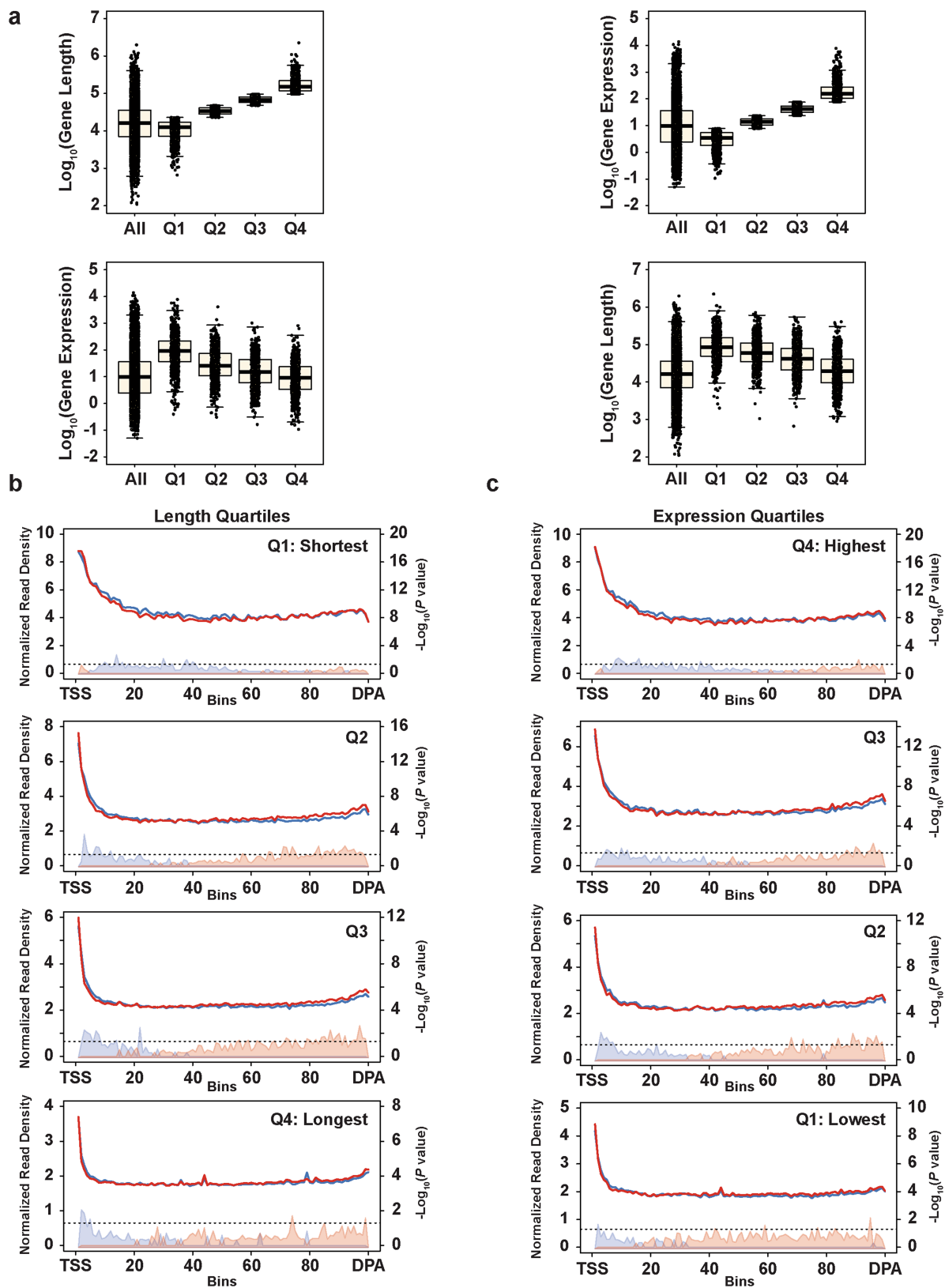
Extended Data Fig. 3 | ChIP antibodies recognizing the same target protein exhibit strongly overlapping metagene patterns. Metagene profiles broken down by individual antibodies used. Blue lines: normalized read density for the specific ChIP antibody in $n = 2$ biological replicates. Orange lines: negative control (combined whole-cell extract and all antibody negative controls, $n = 4$ biological replicates). Black dashed lines:

fold-enrichment (specific ChIP/negative control). Shaded areas: $-\log_{10}$ (bin-wise P values, Kolmogorov–Smirnov one-sided test) of the difference in read depth, with blue shading indicating that the plus CDK12 signal is significantly greater. The $-\log_{10}$ of the P value is shown in the axis on the right, and the horizontal dashed line is $P = 0.05$. TSS, transcription start site. Distal polyA, distal polyadenylation site.



Extended Data Fig. 4 | Schematic of ChIP experiments and data analysis. **a**, Schematic of biological replicate and antibody replicate experimental design. Each ChIP set (RNAPII and Ser2p, in CDK12⁺ or CDK12⁻ cells) consisted of two biological replicates, each ChIP'd with two different antibodies recognizing the same protein. These four

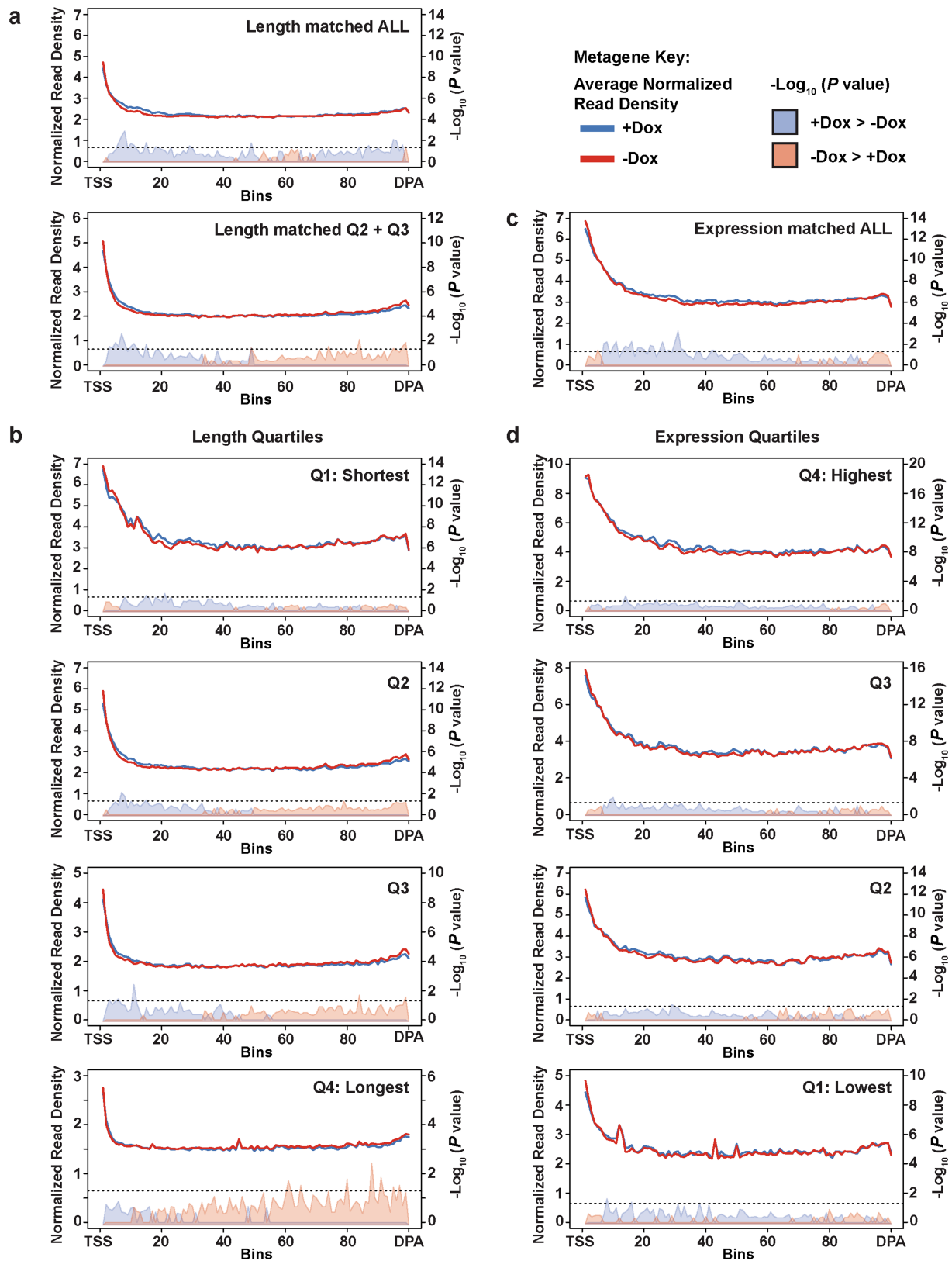
replicates were then combined for the ChIP metagene analyses. **b–d**, Schematic of the steps used to determine average read densities for the ChIP assays, and the statistical test used to determine significant differences in read density that depend on CDK12 expression.



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | RNAPII metagene patterns are influenced by gene length and expression. a, Length and expression quartiles of genes that showed significant changes ($P_{\text{adj}} < 0.05$, FDR adjusted P value determined by the DEXSeq package in R) in IPA or distal isoforms. Boxplots: median value, 25th and 75th quartiles; whiskers show $1.5 \times$ interquartile range. $n = 4$ biological replicates per condition. Top panels: size distributions (\log_{10} of length in nucleotides) of each length quartile (left) and gene expression distributions (\log_{10} of transcripts per million) of each expression quartile (right) compared to the respective distributions of all expressed genes. Bottom panels: expression distributions for each length quartile (left) and length distributions for each expression quartile (right). Note that gene length is generally inversely correlated with expression level, but the median expression of all quartiles of the significantly changing IPA/distal isoforms is higher than the median for all expressed genes. In addition, the median length of all expression quartiles of the significantly changing IPA/distal isoforms is longer than the median for all expressed genes. Thus, the genes that comprise the significantly changing IPA/distal isoform set are longer and more highly expressed for

their length than the broader gene population. **b,** Metagene profiles of RNAPII density in genes with a statistically significant CDK12-sensitive IPA or terminal site divided into length-based quartiles. In the shortest quartile, the CDK12-depleted cells show a trend towards increased density at the 3' end, but the shortest genes terminate before the polymerase can reach a higher density than the CDK12 competent cells. Conversely, the longest genes are expressed at a lower level (see **a**), resulting in a lower RNAPII ChIP signal. For these reasons, the shortest and longest length quartiles were excluded in Fig. 3b, d. **c,** Metagene profiles of RNAPII density in genes with a statistically significant CDK12-sensitive IPA or terminal site divided into expression-based quartiles. **b, c,** Solid lines indicate normalized read density with (blue, $n = 4$ independent ChIPs) or without (red, $n = 4$ independent ChIPs) CDK12; shaded areas indicate $-\log_{10}$ (bin-wise P value, Kolmogorov-Smirnov one-sided test) of the difference in read density (blue indicates that CDK12⁺ signal is greater, pink indicates that CDK12⁻ signal is greater). Horizontal dashed line: $P = 0.05$. DPA, distal polyadenylation site.

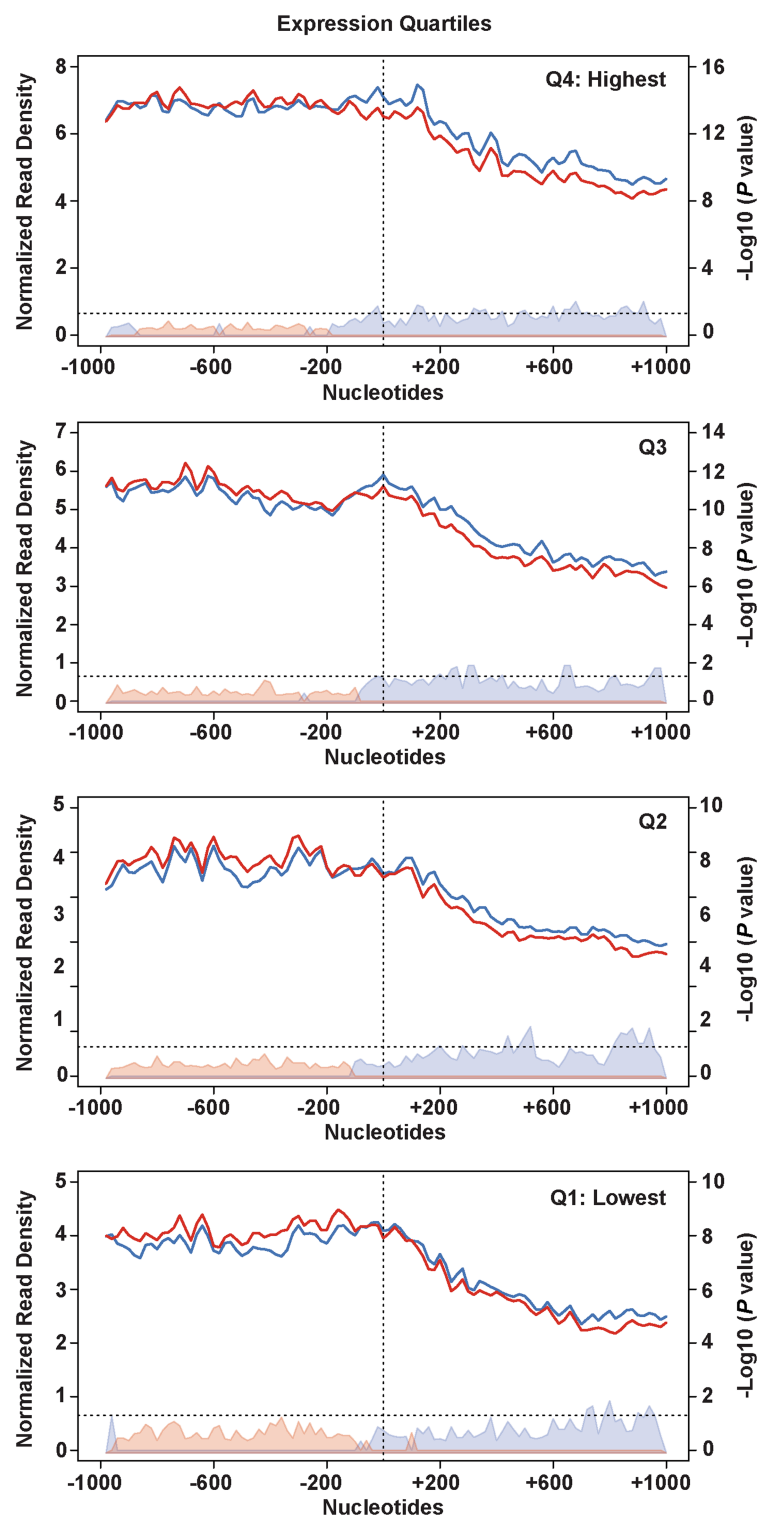


Extended Data Fig. 6 | RNAPII ChIP pattern is not specific to genes with CDK12-sensitive IPAs. **a**, Metagenes profile of RNAPII density in a set of control genes length-matched to the set of genes with significantly changing IPA or distal isoforms. Top, all control genes. Bottom, shortest and longest quartiles removed (as in Fig. 3b). **b**, Metagenes profile of RNAPII density in a set of control genes length-matched to the set of genes with significantly changing IPA or distal isoforms, divided into length quartiles. **c**, Metagenes profile of RNAPII density in a set of control genes expression-matched to the set of genes with significantly changing IPA or

distal isoforms. **d**, Metagenes profile of RNAPII density in a set of control genes expression-matched to the set of genes with significantly changing IPA or distal isoforms, divided into expression quartiles. **a–d**, Solid lines indicate normalized read density with (blue, $n = 4$ independent ChIPs) or without (red, $n = 4$ independent ChIPs) CDK12; shaded areas indicate $-\log_{10}$ (bin-wise P value, Kolmogorov–Smirnov one-sided test) of the difference in read density (blue indicates that CDK12⁺ signal is greater, pink indicates that CDK12[−] signal is greater). Horizontal dashed line: $P = 0.05$.

Metagene Key:
Average Normalized
Read Density
— +Dox
— -Dox

-Log₁₀ (P value)
■ +Dox > -Dox
■ -Dox > +Dox



Extended Data Fig. 7 | Increased RNAPII upstream and decreased RNAPII downstream of first stable nucleosome occurs in all gene expression quartiles. Total RNAPII metagene density 1 kb upstream and 1 kb downstream of the first stable nucleosome for genes with significantly changing IPA or distal isoforms, divided into gene expression quartiles. As in Fig. 3c, solid lines indicate normalized read depth with (blue) or without (red) CDK12, and shaded areas indicate $-\log_{10}$ (bin-wise P values,

Kolmogorov–Smirnov one-sided test) of the difference in read depth, with blue shading indicating that the plus CDK12 signal is significantly greater, and pink shading indicating that the minus CDK12 signal is significantly greater. Horizontal dashed line is $P = 0.05$. Vertical dashed line indicates the position of the first stable nucleosome dyad. $n = 4$ biological replicates per condition.

Metagene Key:

Average Normalized
Read Density

+Dox

-Dox

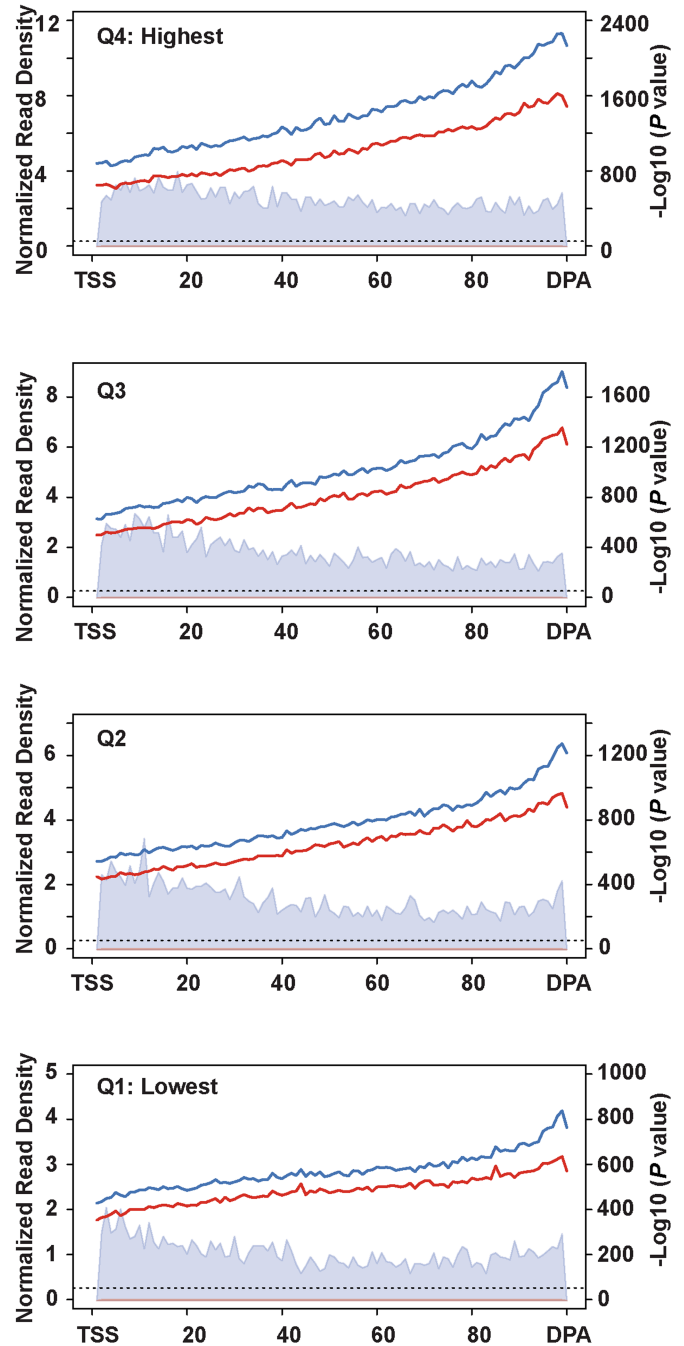
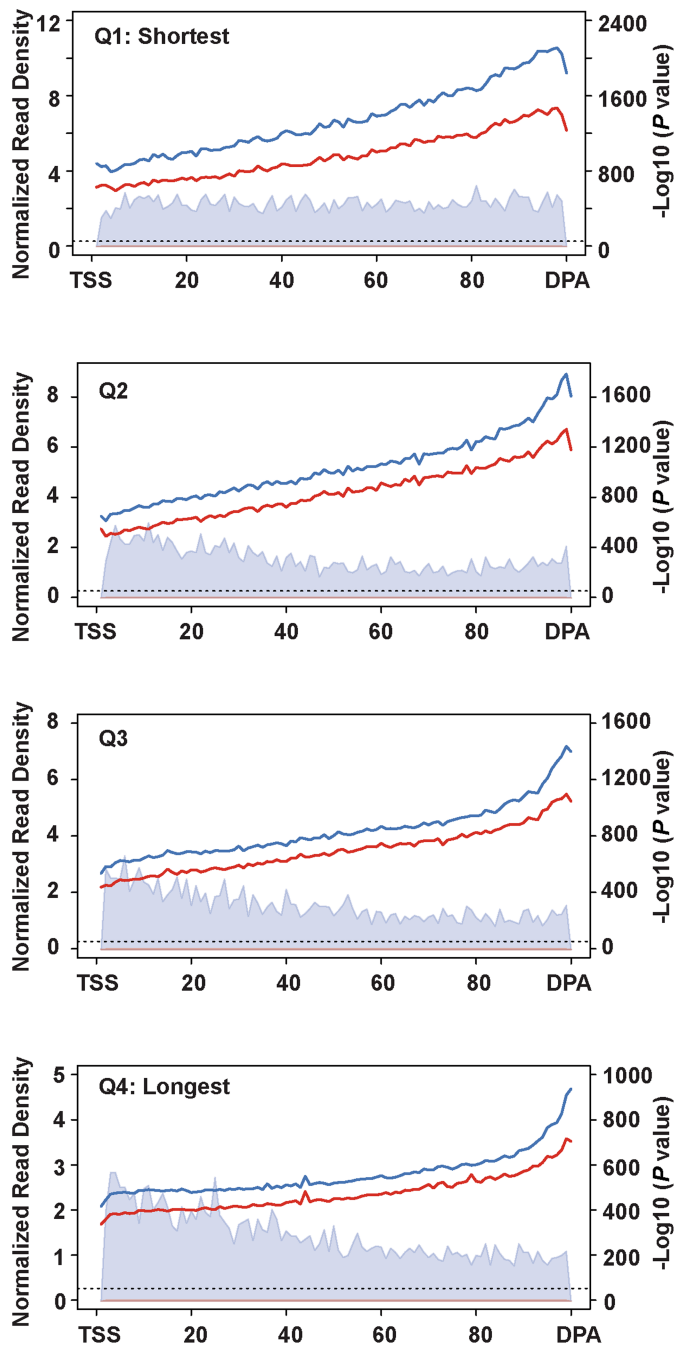
 $-\log_{10}(P \text{ value})$

+Dox > -Dox

-Dox > +Dox

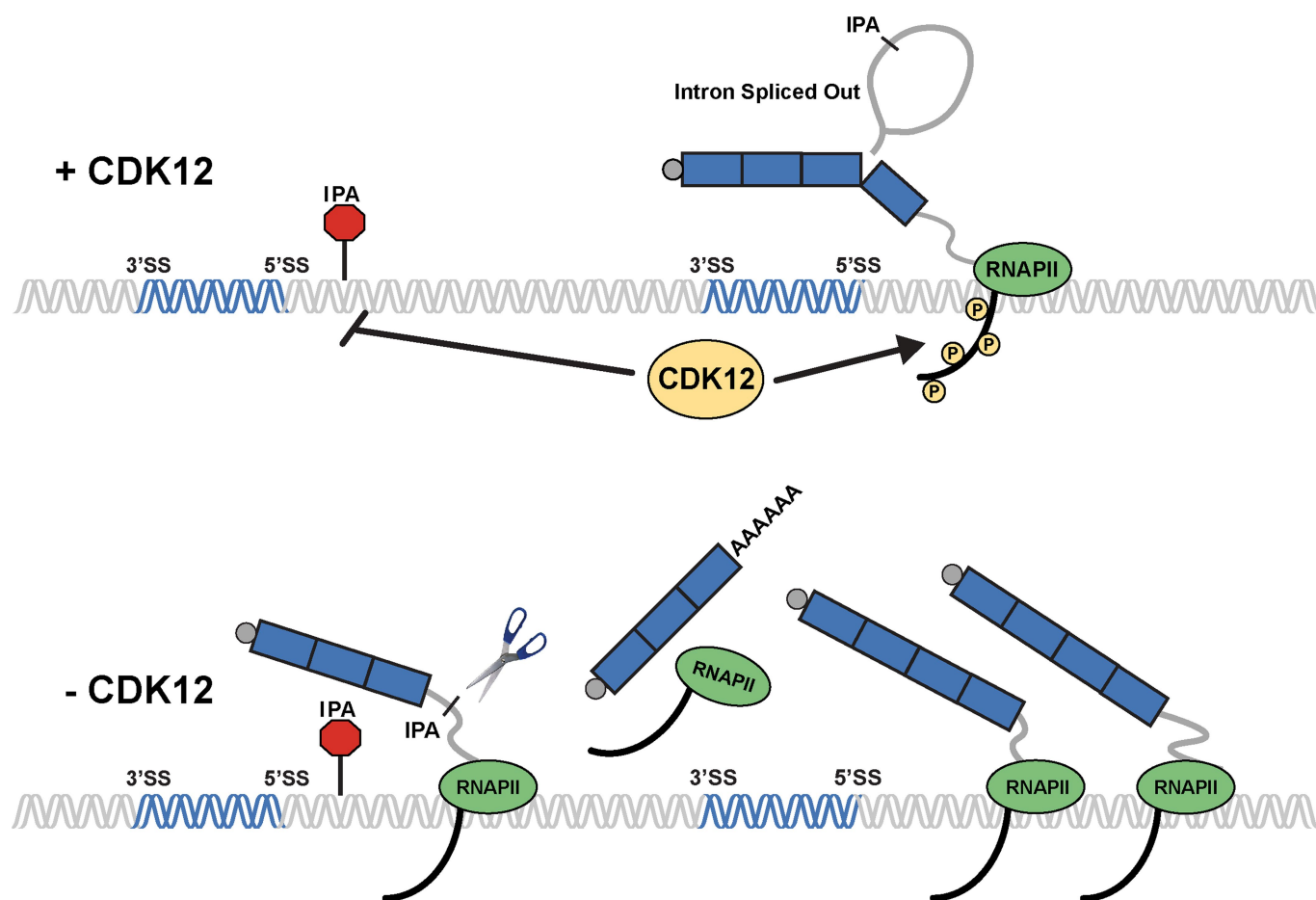
Length Quartiles

Expression Quartiles



Extended Data Fig. 8 | Ser2p is depleted by CDK12 loss and metagene patterns are influenced by gene expression and length. Left, metagene profiles of Ser2p RNAPII density in genes with a statistically significant ($P_{\text{adj}} < 0.05$, FDR adjusted P value determined by the DEXSeq package in R) CDK12-sensitive IPA or terminal site divided into length-based quartiles. As in Fig. 3d, solid blue lines indicate average normalized read

density in CDK12⁺ cells, red solid lines are the average normalized read density in CDK12-depleted samples. Light blue shading indicates that the plus CDK12 signal is significantly greater. $n = 4$ biological replicates per condition. Right, metagene profiles of Ser2p RNAPII density in genes with a statistically significant CDK12-sensitive IPA or terminal site divided into expression-based quartiles. $n = 4$ biological replicates per condition.



Extended Data Fig. 9 | Model for CDK12-dependent effects on gene expression. Top, as RNAPII transcribes through a region of a gene (exonic regions shown in blue with 5' and 3' splice sites (SS) indicated, introns in grey) containing an IPA site (red octagon), CDK12-dependent RNAPII-CTD Ser2 phosphorylation suppresses IPA site usage. Bottom, in the absence of CDK12, RNAPII-CTD Ser2 phosphorylation is decreased. IPA site usage increases, resulting in increased truncated isoforms and decreased distal-most isoforms. RNAPII that transcribes through the downstream exon accumulates with increasing density towards the 3' end of the gene. IPA usage is in competition with the splicing of its

encompassing intron. Decreasing the efficiency of splicing or increasing the activity of cleavage and polyadenylation could both increase IPA usage. Alternatively, a decrease in the efficiency of transcription elongation could alter the kinetic balance to favour IPA usage. Indeed, previous studies have suggested that slower RNAPII elongation rates, due to mutant polymerases or alterations in transcription elongation factors, increase IPA usage over that of distal sites^{48–51}. All three of these possibilities have been related to RNAPII Ser2p, but it is unclear how CDK12-dependent phosphorylation of Ser2p is related to these non-mutually-exclusive possibilities.

Extended Data Fig. 10 | Upregulated IPA usage in human tumours is specific to *CDK12* LOF mutations and not mutations in other BRCAness genes; treatment of human ovarian and prostate cancer cell lines with THZ531 phenocopies the increased IPA site usage observed upon *CDK12* genetic loss. **a**, RNA-seq read density from TCGA tumours from patients with prostate adenocarcinoma or ovarian cystadenocarcinoma with the indicated mutational status at a *CDK12*-sensitive IPA site in the human *ATM* locus (*ATM* IPA #2). Tumours shown in blue are wild-type for *CDK12* and diploid unless marked as amplified (A). Tumours shown in red carry missense putative driver mutations, truncating mutations, or shallow (SD) or deep (DD) gene deletions at the *CDK12* locus. Of note, all of the ovarian cystadenocarcinoma tumours that carry *CDK12* point mutations also have a shallow deletion at the *CDK12* locus except for the tumour with the *CDK12*(R882L) missense mutation, which is diploid across the locus. The 23 tumours in orange harbour putative driver mutations in the other BRCAness genes (*ATM*, *BRCA1*, *BRCA2*, *FANCA*, or *CHEK2*) as noted. **b**, Quantification of usage of two different IPA sites in human *ATM* and at IPA sites in *FANCD2* and *WRN* in human prostate and ovarian tumours from TCGA data (combined in this analysis). Tumours with wild-type or amplified *CDK12* are shown in blue (WT), those with *CDK12* deletions, missense mutations, or truncating

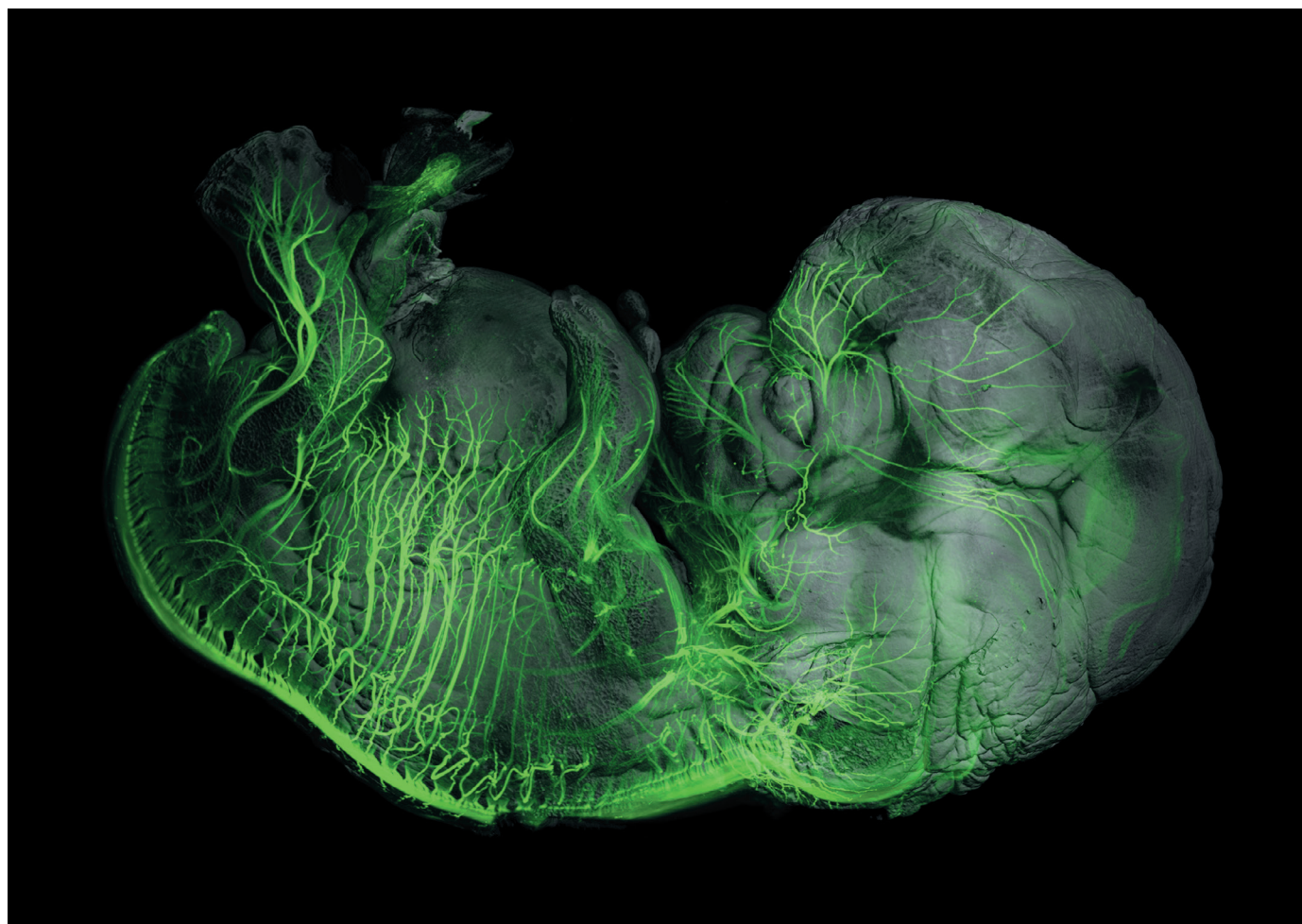
mutations in red (Mut), and those with putative driver mutations in the five BRCAness genes (*ATM*, *BRCA1*, *BRCA2*, *FANCA*, and *CHEK2*) in orange. Medians are indicated by horizontal black bars and sample sizes are indicated below. *P* values were determined by one-sided Mann–Whitney *U* test. **c**, Immunoblots showing the effect of 4 h of THZ531 treatment versus DMSO on RNAPII pSer2 (3E10 antibody) in two prostate carcinoma cell lines (22RV1 and PC-3) and one high-grade serous ovarian carcinoma cell line (OVCAR4). Total RNAPII (8WG16 antibody), HSP90, and Vinculin are shown as loading controls. **d**, **e**, Isoform-specific RT-qPCR used to assay for the expression of IPA and distal polyadenylation isoforms in two prostate carcinoma cell lines (22RV1 and PC-3) and one high-grade serous ovarian carcinoma cell line (OVCAR4) after 4 h of 400 nM THZ531 treatment compared to vehicle (DMSO). Blue bars (DMSO) and red bars (THZ531) represent mean (\pm s.e.m.) for $n = 3$ biological replicates. **d**, Four IPA sites were assayed. Three IPA sites were identified in the TCGA data from human ovarian and prostate tumours (*ATM* IPA #1, *FANCD2* IPA, and *WRN* IPA; Fig. 4f, g and Extended Data Fig. 10a, b). One IPA site corresponded to a significantly changing IPA site in our mES cell *Cdk12* Δ clones (APAF1). **e**, Distal polyadenylation isoforms for the genes in **d**.

TECHNOLOGY FEATURE

TRANSPARENT TISSUES BRING BIOLOGY INTO FOCUS

Techniques that render tissues as clear as glass and swell them to several times their original size are giving unprecedented access to the inner workings of biological systems.

ALAIN CHÉDOTAL & MORGANE BELLE



3D image of peripheral nerves (green) in a tissue-cleared 8-week-old human embryo.

BY MICHAEL EISENSTEIN

In March, researchers in Japan mapped the cellular organization of the mouse brain in unprecedented detail.

Systems biologist Hiroki Ueda at the RIKEN Center for Biosystems Dynamics Research in Osaka, Japan, and his team created an atlas of the mouse brain using a technique called CUBIC-X, in which they chemically labelled every cell in the brain, then rendered the organ crystal-clear while

also expanding its size tenfold¹. From there, they used sophisticated imaging techniques to compile a comprehensive 3D neuronal survey — of some 72 million cells in all, Ueda says. The resulting atlas reduces the brain to a compact database of cellular addresses, which the team used to explore changes in various brain regions during development. Moving forward, the atlas could drive deeper explorations of brain structures that control behaviours such as the sleep–wake cycle.

CUBIC-X is just one component in a growing toolbox of such methods, which exploit readily available chemicals to provide researchers with a window not just into the brain, but into virtually every organ in the body. Some are tissue-clearing methods that make opaque tissues transparent, whereas others complement tissue clearing with a proportional size increase that exposes molecular details to conventional microscopy. The choice comes down to the scientific question. There are many ways to achieve similar ►

► ends, and users should investigate the strengths and limitations of different methods before deciding which to use.

MIND READERS

The hunger for tissue-clearing techniques originated with neuroscientists, who were frustrated by their limited ability to trace the snaking routes of axons and dendrites in the brain.

Such studies conventionally involve serial imaging of thin sections of labelled brain tissue, followed by computational reconstruction into 3D. But the process is slow: imaging circuits in a single mouse brain can entail weeks in front of a microscope. And the resulting maps are only as good as the input data. “Most times, you are only sampling a handful of slices, and this is not very efficient for reconstruction,” says Viviana Gradinaru, a neuroscientist at the California Institute of Technology in Pasadena. “And cutting can damage the tissue surface and edges in a way that prevents you from realigning them.”

A better approach would be to make the tissue transparent and then image it intact. But only in the past few decades have molecular reagents, genetic strategies and imaging techniques advanced far enough to make that possible.

When it comes to illuminating the brain's interior, lipids are public enemy number one. As light passing through an aqueous solution encounters a lipid surface, the change in refractive index causes it to bend and scatter. “Think about Jell-O [a jelly]: it's made mainly of proteins and it's translucent,” says Gradinaru. “But if you add cream to the Jell-O, it becomes opaque. That cream is made of lipids.” Cell and organelle membranes are made mainly of lipids, as are the myelin sheaths enveloping axons. Clearing the brain entails eliminating these molecules, while physically stabilizing the molecules that remain behind.

German anatomist Werner Spalteholz first demonstrated a strategy for clearing opaque tissues in 1911, using chemical solvent treatments that eliminated light-scattering biomolecules. But that method was incompatible with today's fluorescent reagents, and damaging to tissue structures.

In 2011, Hans-Ulrich Dodt, a brain-imaging specialist at the Vienna University of Technology; Ali Ertürk, now at the University of Munich, Germany; and Frank Bradke at the German Center for Neurodegenerative Diseases in Bonn described one of the first modern clearing techniques. Called 3DISCO, the method is a spiritual descendant of Spalteholz's protocol, using a gentler cocktail of chemical solvents that dissolve lipids while preserving cellular structures and dehydrating and hardening the specimen into a clear framework that retains the tissue's original structure². “We think that the solvent-based methods are the most reliable in terms of reproducibility and cost — the first time you do it, it works,” says Alain Chédotal, a developmental neuroscientist at the INSERM Vision Institute

in Paris. Solvent-based methods are generally best suited for use with fluorescently tagged antibodies as the reporter molecules, because genetically expressed fluorescent proteins tend to yield a weakened signal or be denatured by such treatments. But a new variant of 3DISCO from Ertürk's team overcomes that problem. vDISCO uses dye-labelled ‘nanobodies’ to boost the signal from fluorescent proteins in solvent-cleared tissues — an approach the team used to clear and image intact mice (see go.nature.com/2tk6hr3).

Another widely used tissue-clearing option is CLARITY, which Gradinaru helped to develop as a graduate student in neuroscientist Karl Deisseroth's lab at Stanford University in California in 2013 (ref. 3). The Deisseroth lab makes extensive use of fluorescent proteins in its neuroscience research, and sought a more ‘naturalistic’ clearing approach that minimized damage to biomolecules of interest. CLARITY uses detergent to eliminate lipids, while reinforcing the tissue infrastructure with

polymers that form a water-based hydrogel. “All of those monomers hold hands with each other and they lock in the proteins as well,” explains Gradinaru, “and then you can come in with that gentle detergent to take the lipids out.” Early versions of CLARITY were technically challenging, requiring an electrical field to actively drive out detergent-encapsulated lipids. But Gradinaru subsequently devised a simpler alternative that perfuses the animal's vasculature with clearing solution to achieve the same effect.

For the CUBIC family of techniques, Ueda developed yet another approach⁴. CUBIC exploits ‘hydrophilic’ chemicals that draw water into the fixed specimen while pushing dissolved lipids out. As with CLARITY, the

method preserves the structure and function of fluorescent proteins while clarifying the sample.

BODY OF EVIDENCE

Clearing methods allow researchers to tackle neuroscience questions that were previously out of reach. Chédotal, for instance, is using 3DISCO to explore the intact wiring of the visual system. “From the eye, you have the output of ganglion cells that project to maybe 30 different parts of the brain,” he says. “How these axons find these different targets is completely unknown.”

But the methods aren't limited to the brain. “Much to our surprise, most of the rodent organs turned transparent within a few days,” says Gradinaru of her work in developing perfusion-based CLARITY. “It resulted in the whole body being cleared except the skin and bones,” she says. When it comes to tissue clearing, bone poses a particular challenge, Gradinaru says, because calcium continues to reflect light even after conventional clearing. But additional treatments can eliminate this problem. Ueda, for instance, has shown that EDTA, a commonly used laboratory chemical, efficiently removes calcium from bone. Gradinaru's team has also developed a bone-clearing CLARITY variant.

Such methods make it possible to perform whole-body imaging of intact specimens, and researchers have applied them to track rare populations of stem cells or tumour metastases, and even to trace the developing vasculature and peripheral nervous system in first-trimester human embryos. As samples get bigger, however, microscopy can become a bottleneck, and researchers must balance research goals with practicality. Light-sheet microscopy is one popular solution. “You're scanning a plane of light through the tissue rather than a point, and that greatly accelerates the imaging,” says Deisseroth. So, too, is the physical shrinkage that accompanies certain clearing methods. In 3DISCO, dehydration can reduce sample size by up to 50%. “That means we can image a whole human embryo in 3D in one shot,” says Chédotal.

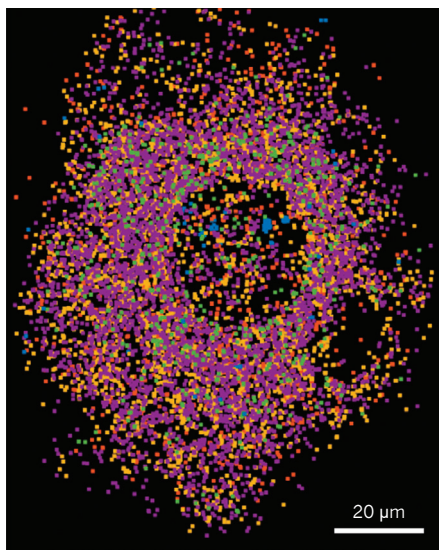
Still, the envelope can be pushed only so far. Chédotal, who has cleared human brains, has had to content himself with imaging regions measuring a few cubic centimetres, roughly 1% of the organ. “I could make a whole cow transparent,” he says. “But I wouldn't be able to image it, so what's the point?”

THE BIG PICTURE

Clearing is a valuable starting point for tissue imaging, but to make out fine molecular details, researchers also need the means to ‘zoom in’.

Conventional light microscopes are unable to distinguish molecules separated by less than the ‘diffraction limit’ of light, around 200 nanometres — a problem that led to the development of technically sophisticated super-resolution microscopy approaches. Ed Boyden, a neurobiologist at the

Expansion microscopy promises super-resolution imaging without the costly hardware.



Expansion microscopy composite image showing six labelled RNA targets in a cultured HeLa cell.

Transparent tumours

Cancer medicine is advancing fast, guided by new insights into immunology and genetics. Yet cancer pathology remains rooted in the past, relying on staining individual slices of tumour tissue for microscopic examination. “It’s very conservative, and the young clinical scientists are a little bit depressed by that situation,” says Hiroki Ueda, a systems biologist at the RIKEN Center for Biosystems Dynamics Research in Osaka, Japan.

Tissue-clearing methods offer an alternative. In 2017, for instance, Ueda and his colleagues showed that the tissue-clearing technique CUBIC confers greater sensitivity than conventional preparation

methods, allowing clinicians to peer into larger sections of tissue and home in on features that might otherwise be overlooked¹⁰. “Sometimes early-stage cancer gets misdiagnosed,” says Ueda. “But if you look at the tissues with clearing in three dimensions, we rarely miss the cancer.”

Tissue-clearing strategies also offer insights into tumour biology. Neurobiologists Karl Deisseroth at Stanford University in California and Per Uhlén at the Karolinska Institute in Stockholm applied a variant of the clearing technique 3DISCO to get an unprecedented view of tumour heterogeneity — a factor that can profoundly affect treatment response¹¹.

And at the Massachusetts Institute of Technology in Cambridge, neurobiologist Ed Boyden and his team have shown that the tissue-clearing and resolution boost from expansion microscopy enables more-accurate diagnosis of neoplastic lesions that can herald early-stage breast cancer¹².

Both Boyden and Deisseroth have founded companies to advance clinical development of their techniques, and some practitioners of tissue clearing already foresee the beginning of the end for conventional methods. “Young pathologists are so excited about our technology,” says Ueda. “It seems to be the future to them.” **M.E.**

Massachusetts Institute of Technology (MIT) in Cambridge, was joking when he initially proposed an alternative approach — ‘blowing up’ the brain. But soon he saw real potential in the idea. “All the proteins in the cell are jam-packed together, and if we expand them apart from each other, maybe we could see them better,” he says.

Boyden’s team drew on decades of research into the properties of hydrogels that undergo proportional expansion on hydration. In their first-generation ‘expansion microscopy’ method⁵, published in 2015, samples are treated with specially designed fluorescent labels recognizing target molecules of interest, and then incubated with a polymer solution that forms a hydrogel matrix. The labels attach to this matrix, locking them into position relative to each other. Finally, the surrounding tissue is broken up through chemical or enzymatic treatment, and hydrated to swell the gel matrix. The resulting expansion leaves the labels in the same relative position, but separated by up to four times their original distance. As a result, molecules that previously were too close together to distinguish can now be discerned using standard fluorescence microscopy.

For many biologists, the method seemed too good to be true. “My first thought was: ‘This is crazy, how could that even work?’” recalls Joshua Vaughan, a bioimaging researcher at the University of Washington in Seattle. But he was intrigued enough to try it — and subsequently devised an alternative version that uses conventional fluorescent proteins or antibodies rather than specially designed labelling reagents. Later variants include an iterative method from Boyden’s group that uses two rounds of treatment to achieve up to 20-fold expansion, and an alternative method developed by Kwanghun Chung, a biomedical engineer at MIT, that denatures biomolecules rather than digesting them, which better protects endogenous proteins and the integrity of tissue structures⁶. “We use this method for connectivity mapping, which requires

preservation of neuronal fibres: once you cut them, you lose information,” says Chung. Crucially, these various methods also induce tissue clearing, allowing users to peer deep inside their super-sized samples.

Vaughan’s team has applied expansion microscopy to specimens ranging from fruit-fly larvae to the human kidney, and other researchers are applying it in the clinic (see ‘Transparent tumours’). But it can take considerable trial and error to ‘tenderize’ a new specimen for expansion. “Human kidney has some tough connective tissue,” says Vaughan, “so we just had to keep trying enzymes to get at it.” It is also essential to make sure that expansion really is ‘isotropic’, or equivalent in all directions. But with care and practice, remarkable uniformity can be achieved. “Our results show that distortion is less than 5%, which is equivalent to, or lower than, the distortion you get during sample-mounting,” says Chung.

GOING DEEPER

For cash-strapped biologists, expansion microscopy promises super-resolution imaging without the costly hardware. “Basically anybody can do this,” says Helge Ewers, a cell biologist at the Free University of Berlin, “and any normal microscope can become super-res capable.” Crucially, the reagents required are not particularly exotic, even if some dabbling is required to identify the right formulation. Chédotal notes that clearing an entire mouse costs roughly a dollar. The equipment requirements are equally modest. “You just put the organ into the chemical, and usually just an incubator and shaker are needed — most labs have such devices,” says Ueda. Expansion is more technically demanding than clearing-only methods, but Boyden and others have published ‘best practice’ protocols to help eliminate guesswork in expansion microscopy⁷. “It’s not quite a ‘cookbook’ yet, but we’re getting there,” Boyden says.

Researchers are beginning to explore the potential rewards of such methods. Several

brain-atlas initiatives are now under development, with the goal of going beyond cellular censuses to chart the interconnections between cells. Others are applying clearing and expansion to DNA and RNA, allowing them to survey gene expression as well as proteins. Biophysicist Xiaowei Zhuang at Harvard University in Cambridge, Massachusetts, devised an expansion-microscopy-based strategy that allowed her team to quantify expression of more than 100 genes at the single-RNA level⁸. In parallel, Deisseroth and colleagues have developed a technique called STARmap, in which up to 1,020 different genes can be directly sequenced within cleared brains⁹.

The resulting data can help to classify individual cells. But it could also be layered atop circuit maps and live-animal experiments to reveal the interplay of structure and function in the brain, exposing previously hidden connections. “Because you’re preserving the 3D arrangement of cells in the tissue, we should be able to register this cellular-resolution transcriptomics and connectomics data with cellular-resolution activity patterns collected in real time from living animals,” says Deisseroth. “Tissue clearing allows you to bring all these fundamentally different data streams together.” ■

Michael Eisenstein is a freelance writer based in Philadelphia, Pennsylvania.

1. Murakami, T. C. *et al. Nature Neurosci.* **21**, 625–637 (2018).
2. Ertürk, A. *et al. Nature Protoc.* **7**, 1983–1995 (2012).
3. Chung, K. *et al. Nature* **497**, 332–337 (2012).
4. Susaki, E. A. *et al. Cell* **157**, 726–739 (2014).
5. Chen F., Tillberg P. W. & Boyden, E. S. *Science* **347**, 543–548 (2015).
6. Murray, E. *et al. Cell* **163**, 1500–1514 (2015).
7. Asano, S. M. *et al. Curr. Protoc. Cell Biol.* **80**, e56 (2018).
8. Wang, G. *et al. Sci. Rep.* **8**, 4847 (2018).
9. Wang, X. *et al. Science* **361**, eaat5691 (2018).
10. Nojima, S. *et al. Sci. Rep.* **7**, 9269 (2017).
11. Tanaka, N. *et al. Nature Biomed. Eng.* **1**, 796–806 (2017).
12. Zhao, Y. *et al. Nature Biotechnol.* **35**, 757–764 (2017).

CAREERS

SHARE Tell us your career story at
naturecareerseditor@nature.com

MENTAL HEALTH Find advice and support at
go.nature.com/wellbeing

INSTAGRAM Follow us at
[instagram.com/naturejobs](https://www.instagram.com/naturejobs)

ANDREAS W. MATTHES



Emiliano Monroy-Ríos samples rocks from a cave off the eastern coast of Mexico with adviser Patricia Beddows.

MENTORSHIP

A chance to grow

Knowing when to hand-hold and when to step back is crucial for helping junior scientists.

Supervisors can help to shape the lives and careers of their students and trainees. Sometimes, they become lifelong mentors and eventual collaborators, contributing to a new generation of scientific discovery. And students can forge meaningful relationships with those senior scientists even at the earliest stages of their science careers.

In *Nature's* 2017 global PhD survey, 34% of respondents said that a supervisor helped them to reach their current career decision (*Nature* 550, 549–552; 2017). Most respondents said that they were happy with their adviser, but nearly 25% said they would switch if they could.

However, supervisors must often learn to lead from their own experiences and mistakes. Here, four researchers at different career levels share stories of good supervisor relationships they have experienced, and what made those relationships so effective.

BRYONY JAMES Identify what works

Materials engineer, University of Auckland, New Zealand

My PhD supervisor, Barry Welch, treated me as a junior colleague, not just as a graduate student. I was part of the team, and so I had to step up. If I didn't know what I was meant to be doing, I was expected to try to work it out for myself before asking for help.

In my first year, I was developing new experimental techniques using high-temperature furnaces to look at the oxidation of carbon. The furnaces were built in-house

and were incredibly temperamental. My supervisor's attitude was: "Well, it's your furnace, figure it out." So I did.

When one of the heating elements broke, I started taking the furnace apart and fixing it. My PhD was not on designing and building furnaces — it was on the oxidation of carbon. But because I had to take the furnace apart, rebuild it and think about how it was operating, I gained a much more intimate understanding of the equipment I was using, which allowed me to understand the context of my results. Taking the ground-up approach of stripping something right down to its basic components and building it again — if you have the time to do it — is a powerful learning experience. None of that is going to end up in your thesis, but it gives you enormous confidence in the results you get from your equipment.

That worked really well for me because ►

► I'm an independent learner. If someone had tried to be more hands-on, supervising every step of my PhD path, I would have found it absolutely claustrophobic. I'm now supervising my own students, and I take a similar approach to that of my PhD supervisor. I expect my students to have a really good go at things before they ask for help.

I've learnt from well over 20 years of supervising PhD students that what worked for me does not necessarily work for everyone. I suspect that my supervision style works only for people who want independence. So, when I interview prospective PhD students, I am clear about that. Several times I have suggested that potential PhD candidates speak to one colleague or another across the university where I sense there would be a better outcome for the student.

You have to be very honest with yourself and your students. Some people want to be told what to do. I generally don't end up supervising those students. You have to have a good match between supervisor and student.

EMILIANO MONROY-RÍOS

She believed in me

PhD candidate in hydrogeology, Northwestern University, Evanston, Illinois

I was born in Mexico City. My father always took me to the ocean on summer trips, and I wanted to be an oceanographer as a child. I went on to study chemistry as an undergraduate student in Mexico City, did a master's degree in limnology and then moved to the Riviera Maya on the Yucatán Peninsula. That region has the longest underwater caves in the world.

As a research assistant, I met my current PhD adviser, Patricia Beddows. I helped her by entering and mapping a dry cave, and then became a technical cave diver. She examined the side projects I was working on at the time and pushed me to apply for fellowships and scholarships to pursue a PhD in her lab at Northwestern University.

I had always thought about doing a PhD. But after my master's, as a research assistant, I felt stuck. Beddows believed in me and got me thinking about a PhD again. It was like a revival in lost confidence.

I moved to Chicago in January 2011. I had been living in a tropical paradise, and the next day, I was walking in the snow, thinking, "What am I doing here?" Patricia and her husband Edward gave me tips for surviving winter in the city. But I was depressed and seriously thought two or three times about quitting. I was feeling really, really bad and was afraid that my academic performance was declining. I thought, "I cannot make it."

Patricia understood that my health came first,

before my research. She stated very clearly that if I made the decision to quit, she would support me. But she also convinced me that my work was worth it. It was important for me to have that balance of, "OK, I understand, and I believe you — that you are passing through a bad time. But your work deserves fighting for." So I took a quarter off to go back to Mexico for the winter of 2013. We talked while I was gone. If it weren't for her, I would have quit.

ADRIANE LAM

Give us chances to grow

PhD candidate in geosciences, University of Massachusetts, Amherst

Neither of my parents went to college. When I transferred to James Madison University in Harrisonburg, Virginia, from a two-year college as an undergraduate, I was shy and uncertain and didn't feel comfortable. I took a palaeoclimatology course with Kristen St. John, who became my undergraduate research adviser. We met weekly. She really showed me the ropes — she told me how to find published studies and how to interpret them. She taught me the importance of networking and how to collaborate early in my career. That came in handy later.

One afternoon, I had been working for many hours in the lab. She came in and said, "The level at which you're working is like a master's student." I had really wanted to do a master's degree, but never thought that I could. But when she said that, I thought, maybe I can.

During my master's-degree programme, my supervisor, Alycia Stigall, knew when to push me and when to leave me alone. I was doing a lot of modelling and I always had coding problems. I would complain to Alycia, and she would say: "You can figure it out." And I always would. I needed a push at that point.

I have published several papers with Alycia, including one with two other alumni of the Stigall lab. Each time, we had an open and honest conversation about the authorship list. She showed me that advisers should have conversations about the authorship protocol when they publish with students, because author order can cause contention among lab members.

Alycia also taught me the value of working with the public, and we did a lot of volunteering together. We visited grade-school students, went to a US national forest to chat about fossils with the public and ran a workshop with teachers at the Cincinnati Museum Center in Ohio. My outreach experiences led me to realize that I love teaching geological concepts to people of all ages and backgrounds.

Supervisors, push your students. Many of us are afraid to step outside our comfort zones. Give us opportunities to grow as scientists.

HANNAH REICH

They pushed me beyond their lab

PhD candidate in biology, Pennsylvania State University, University Park

During the final year of my undergraduate studies at Clark University in Worcester, Massachusetts, my adviser, Deborah Robertson, helped me to pioneer a collaboration with Gretchen Goodbody-Gringley at the Bermuda Institute of Ocean Sciences in St. George's. This collaboration allowed me to conduct research on juvenile coral in Bermuda and to transport samples back to Clark, where I completed molecular lab work for my master's degree.

These researchers hadn't been working together before. Being able to connect them and have my own niche was exciting to me, and the collaborative and explorative approach to science championed by Deborah and Gretchen is something I have continued to follow during my doctoral research. My supervisor, Todd Lajeunesse, unhesitatingly let me spend a couple of summers in Taiwan with oceanographers and chemists to conduct my PhD research.

Mentors have pushed me beyond their labs. In an ideal situation, having a mixture of mentors with different academic strengths, cultural backgrounds and advising styles allows the student to observe and internalize their mentoring expertise and what makes each of them a great scientist. Because I had shifted towards being more globally networked, I knocked on doors that weren't necessarily labelled as open, seeking collaboration and idea exchanges.

Ultimately, excellent mentorship boils down to thinking ahead and supporting students, especially when they are exploring uncharted waters. I am drawn to mentors who encourage a student to have many mentors. Principal investigators have access to different inner circles of people they frequently interact with in academia, where various opportunities are often shared and discussed. Deborah, for example, talked up my work to Clark's media office, which led to this interview. That just goes to show that the best advisers advocate for their students in situations that they wouldn't necessarily be involved in, or even aware of. ■

INTERVIEWS BY EMILY SOHN

These interviews have been edited for clarity and length.

CORRECTION

The Spotlight article 'Science in Colombia on the cusp of change' (*Nature* **562**, S109–S111; 2018) erroneously stated that Colombia is bordered by the Atlantic Ocean. In fact, it is the Pacific Ocean.

A BEGINNER'S GUIDE TO SPACE TRAVEL AND SEAFOOD

Are you ready for a new life?

BY STEVEN FISCHER

You never want to be on the first arkship.

They won't tell you that at the Travel Bureau or in those holos the Colony Department spews out, but it's good advice. You can trust me.

Listen, I get it. I was young once, too. All spitfire and stardust and dreams of going someplace new. Wide-eyed at the first travel agent who rolled by the arcology in a shiny new hover skid.

They're slick, those recruiters. Reel you in with pretty pictures of worlds that haven't been bulldozed and butchered. Planets that still boast fresh water and trees. Places you can stretch out your arms without bumping into another carbon breather. Worlds where you can find that special someone and snag yourself a place in the hills. Maybe pop out an offspring or two. Hell, pop out as many as you'd like; there aren't population restrictions yet. I get it, like I said.

Plus, the flight won't be so bad. The trip's 1,000 years, but you'll be asleep almost the whole time, comfortable as a chefbot in a kitchen, nestled into your own little fugue. Just pull the lid down, pump some cryo in your veins, and wake up once or twice for your semi-centennial health check. Then, a few quick naps later, you wake up for good on your own personal Eden, right?

Wrong.

Here's what they don't tell you. Two hundred years into that little nap of yours, another arkship leaves for your new home sweet home. Only this one was built two centuries after yours. All the newest gizmos and gadgets. Fugue chambers with double the legroom. Health scanners so efficient you can forget the old turn-and-cough at your next check-up. A gravity drive that's twice as fast as yours.

What's that? Twice as fast? But didn't the agent tell you that your ship was state of

the art? That the principles of physics suggested — hell, demanded — it would be the fastest piece of titanium in the galaxy for 5,000 years? Well, there were some folks a few millennia ago who thought they'd built a boat that couldn't sink. I bet you can guess how that one turned out.

What that means is by the time your old clunker arrives on planet Whatevercrapthey-named it, ship number two has already been there for 300 years. Instead of waking up as an intrepid pioneer, you're going to be serving nutrient smoothies for minimum wage to ship two's great-great-grandkids.

NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

What's that? You were an engineer? That's nice. Too bad

your degree is a millennium out of date.

Oh, and by the way, ship three left 100 years after ship two and is scheduled to arrive next, so you might as well forget that smoothie-stand job; they've got 300 years of education on you.

Best case, maybe you'll scrounge up the cash to buy yourself a little vendor cart. Sell some paella to the folks who were smarter than you, and the tourists who were even smarter than them. Because if there's one thing they haven't been able to improve on, it's a good pan of paella, and for all the lies that recruiter told you, planet Horriblyobviousattemptto-invokenostalgiaforadead-earth does have some damn fine shrimp. And maybe, just maybe you'll have the chance to chat up some young kid and stop them from making the same mistake you did.

So trust me. Put the bowl down, quit stuffing your face, and toss that ticket in your hand straight into the trash. Even better, burn it.

I know. You didn't come this far just to give up, and some old codger's words aren't going to be enough to erase those pictures of Perfectlyfocusgroupedtoappealtothelargestdemographic. I get it. Well, today's your lucky day.

See that hover skid across the street? The guy inside is a good friend of mine. Runs a little transportation service of his own. Newest tech, fastest ships, guaranteed. Don't even run on a gravity drive. One hundred per cent singularity propulsion. They can get you where you're going in a quarter of the time. That's only 250 years.

And let me guarantee you, there is no way anyone's going to come up with something faster by then. ■

Steven Fischer is a medical resident living in the Pacific Northwest. When he's not cracking open a textbook (or a patient's thorax), he can be found exploring the Cascades by bike, boat or boot. You can read more of his work at www.stevenbfischer.com.

ILLUSTRATION BY JACEY

